Approximate factor analysis model building via alternating I-divergence minimization

Lorenzo Finesso and Peter Spreij

Lorenzo Finesso
Institute of Biomedical Engineering
ISIB-CNR
Padova
Italy
e-mail: lorenzo.finesso@isib.cnr.it

Peter Spreij
Korteweg-de Vries Institute for Mathematics
Universiteit van Amsterdam
Amsterdam
The Netherlands
e-mail: spreij@uva.nl

Abstract: Given a positive definite covariance matrix $\widehat{\Sigma}$, we strive to construct an optimal *approximate* factor analysis model $HH^\top + D$, with H having a prescribed number of columns and $D > 0$ diagonal. The optimality criterion we minimize is the I-divergence between the corresponding normal laws. Lifting the problem into a properly chosen larger space enables us to derive an alternating minimization algorithm à la Csiszár-Tusnády for the construction of the best approximation. The convergence properties of the algorithm are studied, with special attention given to the case where D is singular.

AMS 2000 subject classifications: Primary 62H25; secondary 62B10.
Keywords and phrases: approximate factor analysis, I-divergence, alternating minimization.

1. Introduction

Factor analysis (FA), in its original formulation, deals with the linear statistical model

$$Y = HX + \varepsilon, \tag{1.1}$$

where H is a deterministic matrix, X and ε are independent random vectors, the first with dimension smaller than Y, the second with independent components. What makes this model attractive in applied research is the *data reduction* mechanism built in it. A large number of observed variables Y are explained in terms of a small number of unobserved (latent) variables X perturbed by the independent noise ε. Under normality assumptions, which are the rule in the standard theory, all the laws of the model are specified by covariance matrices. More precisely, assume that X and ε are zero mean independent normal vectors

with $\mathbb{C}\mathrm{ov}(X) = P$ and $\mathbb{C}\mathrm{ov}(\varepsilon) = D$, where D is diagonal. It follows from (1.1) that $\mathbb{C}\mathrm{ov}(Y) = HPH^\top + D$. Since in the present paper, basically only covariances are considered, the results obtained will also be valid, in a weaker sense, in a non Gaussian environment.

Building a factor analysis model of the observed variables requires the solution of a difficult algebraic problem. Given $\widehat{\Sigma}$, the covariance matrix of Y, find the triples (H, P, D) such that $\widehat{\Sigma} = HPH^\top + D$. As it turns out, the right tools to deal with the construction of an exact FA model come from the theory of stochastic realization, see Finesso and Picci (1984) for an early contribution on the subject. Due to the structural constraint on D, assumed to be diagonal, the existence and uniqueness of a FA model are not guaranteed.

In the present paper we strive to construct an optimal *approximate* FA model. The criterion chosen to evaluate the closeness of covariances is the I-divergence between the corresponding normal laws. We propose an algorithm for the construction of the optimal approximation, inspired by the alternating minimization procedure of Csiszár and Tusnády (1984) and Finesso and Spreij (2006).

The remainder of the paper is organized as follows. The FA model is introduced in Section 2 and the approximation problem is posed and discussed in Section 3. Section 4 recasts the problem as a double minimization in a larger space, making it amenable to a solution in terms of alternating minimization. It will be seen that both resulting I-divergence minimization problems satisfy the so-called Pythagorean rule, guaranteeing the optimality. In Section 5, we present the alternating minimization algorithm, provide alternative versions of it, and study its asymptotical properties. We also point out, in Section 6, the relations and differences between our algorithm and the EM-algorithm for the estimation of the parameters of a factor analysis model. Section 7 is dedicated to a constrained version of the optimization problem (the singular D case) and the pertinent alternating minimization algorithm. The study of the singular case also sheds light on the boundary limit points of the algorithm presented in Section 5. In the Appendix we have collected some known properties on matrix inversion and I-divergence between normal distributions for easy reference, as well as most proofs of the technical results.

The present paper is a considerably extended version of Finesso and Spreij (2007), moreover providing easier proofs of some of the results already contained in that reference.

2. The model

Consider independent random vectors Z and ε, of respective dimensions k and n, both normally distributed with zero mean. For simplicity $P = \mathbb{C}\mathrm{ov}(Z)$ is assumed to be invertible. For any $n \times k$ matrix L let the random vector Y, of dimension n, be defined by

$$Y = LZ + \varepsilon. \tag{2.1}$$

The linear model (2.1), ubiquitous in Statistics, becomes the standard Factor Analysis (FA) model under the extra constraints

$$k < n, \quad \text{and} \quad \mathbb{C}\text{ov}(\varepsilon) = D \geq 0, \ \text{diagonal}.$$

In many applications one starts with a given, zero mean normal vector Y, and wants to find the parameters P, L, and D of a FA model for Y. The above constraints impose a special structure to the covariance of Y,

$$\mathbb{C}\text{ov}(Y) = LPL^\top + D, \tag{2.2}$$

which is non generic since $k < n$ and D is diagonal, therefore not all normal vectors Y admit a FA model. To elucidate this, consider the joint normal vector

$$V = \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} L & I \\ I & 0 \end{pmatrix} \begin{pmatrix} Z \\ \varepsilon \end{pmatrix}, \tag{2.3}$$

whose covariance matrix is given by

$$\mathbb{C}\text{ov}(V) = \begin{pmatrix} LPL^\top + D & LP \\ PL^\top & P \end{pmatrix}. \tag{2.4}$$

The constraints imposed on $\mathbb{C}\text{ov}(V)$ by the FA model are related to a conditional independence property.

Lemma 2.1. *Let $Y \in \mathbb{R}^n$ be a zero mean normal vector, then $\mathbb{C}\text{ov}(Y) = LPL^\top + D$, for some (L, P, D), with $L \in \mathbb{R}^{n\times k}$, $P > 0$, and diagonal $D \geq 0$ if and only if there exists a k-dimensional zero mean normal vector Z, with $\mathbb{C}\text{ov}(Z) = P$, such that the components of Y are conditionally independent given Z.*

Proof. Assume that $\mathbb{C}\text{ov}(Y) = LPL^\top + D$ and construct a matrix Σ as in the right hand side of (2.4). Clearly $\Sigma \geq 0$, since $P > 0$ and $D \geq 0$, and therefore it is a bonafide covariance matrix, hence there exists a multivariate normal vector V whose covariance matrix is Σ. Writing $V^\top = (Y^\top, Z^\top)^\top$ for this vector, it holds that $\mathbb{C}\text{ov}(Z) = P$, moreover $\mathbb{C}\text{ov}(Y|Z) = D$ (see Equation (A.1)). The conditional independence follows, since D is diagonal by assumption. For the converse assume there exists a random vector Z as prescribed in the Lemma. Then $\mathbb{C}\text{ov}(Y|Z)$ is diagonal by the assumed conditional independence, while $E(Y|Z) = LZ$ for some L, being a linear function of Z. We conclude that $\mathbb{C}\text{ov}(Y) = \mathbb{C}\text{ov}(E(Y|Z)) + \mathbb{C}\text{ov}(Y|Z) = LPL^\top + D$ as requested. □

The above setup is standard in system identification, see Finesso and Picci (1984). It is often convenient to give an equivalent reparametrization of model (2.3) as follows. Let $P = Q^\top Q$, where Q is a $k \times k$ square root of P, and define $X = Q^{-\top} Z$. Model (2.3) then becomes

$$V = \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} LQ^\top & I \\ Q^\top & 0 \end{pmatrix} \begin{pmatrix} X \\ \varepsilon \end{pmatrix},$$

where $\mathbb{C}\mathrm{ov}(X) = I$. The free parameters are now $H = LQ^\top$, the diagonal $D \geq 0$, and the invertible $k \times k$ matrix Q. In this paper we will mostly, but not always, use the latter parametrization, which will be written directly in terms of the newly defined parameters as

$$V = \begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} H & I \\ Q^\top & 0 \end{pmatrix} \begin{pmatrix} X \\ \varepsilon \end{pmatrix}, \tag{2.5}$$

for which

$$\mathbb{C}\mathrm{ov}(V) = \begin{pmatrix} HH^\top + D & HQ \\ (HQ)^\top & Q^\top Q \end{pmatrix}, \tag{2.6}$$

Note that, with this parametrization,

$$Y = HX + \varepsilon, \qquad \text{and} \quad \mathbb{C}\mathrm{ov}(Y) = HH^\top + D. \tag{2.7}$$

For simplicity, in the first part of the paper, it will be assumed that H has full column rank and $D > 0$.

3. Problem statement

Let Y be an n dimensional, normal vector, with zero mean and $\widehat{\Sigma} = \mathbb{C}\mathrm{ov}(Y)$ given. As a consequence of Lemma 2.1 it is not always possible to find an exact FA analysis model (2.3), nor equivalently (2.5), for Y. As it will be proved below, one can always find a best approximate FA model. Here 'best' refers to optimizing a given criterion of closeness. In this paper we opt for minimizing the I-divergence (*a.k.a.* Kullback-Leibler divergence). Recall that, for given probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$, defined on the same measurable space, and such that $\mathbb{P}_1 \ll \mathbb{P}_2$, the I-divergence is defined as

$$\mathcal{I}(\mathbb{P}_1||\mathbb{P}_2) = \mathbb{E}_{\mathbb{P}_1} \log \frac{\mathrm{d}\mathbb{P}_1}{\mathrm{d}\mathbb{P}_2}. \tag{3.1}$$

In the case of normal laws the I-divergence (3.1) can be explicitly computed. Let ν_1 and ν_2 be two normal distributions on $\mathbb{R}^m$, both with zero mean, and whose covariance matrices, Σ_1 and Σ_2 respectively, are both non-singular. Then the distributions are equivalent and the I-divergence $\mathcal{I}(\nu_1||\nu_2)$ takes the explicit form, see Appendix A,

$$\mathcal{I}(\nu_1||\nu_2) = \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{m}{2} + \frac{1}{2} \mathrm{tr}(\Sigma_2^{-1}\Sigma_1). \tag{3.2}$$

Since, because of zero means, the I-divergence only depends on the covariance matrices, we usually write $\mathcal{I}(\Sigma_1||\Sigma_2)$ instead of $\mathcal{I}(\nu_1||\nu_2)$. Note that $\mathcal{I}(\Sigma_1||\Sigma_2)$, computed as in (3.2), can be considered as a I-divergence between two positive definite matrices, without referring to normal distributions. Hence the approximation Problem 3.1 below, is meaningful also without normality assumptions.

The problem of constructing an approximate FA model, i.e. of approximating a given covariance $\widehat{\Sigma} \in \mathbb{R}^{n \times n}$ by $HH^\top + D$, can be cast as the following

Problem 3.1. Given $\widehat{\Sigma} > 0$ of size $n \times n$ and an integer $k < n$, minimize

$$\mathcal{I}(\widehat{\Sigma}||HH^\top + D) = \frac{1}{2}\log\frac{|HH^\top + D|}{|\widehat{\Sigma}|} - \frac{n}{2} + \frac{1}{2}\mathrm{tr}((HH^\top + D)^{-1}\widehat{\Sigma}), \tag{3.3}$$

where the minimum, if it exists, is taken over all diagonal $D \geq 0$, and $H \in \mathbb{R}^{n\times k}$.

Note that $\mathcal{I}(\widehat{\Sigma}||HH^\top + D) < \infty$ if and only if $HH^\top + D$ is invertible, which will be a standing assumption in all that follows.

The first result is that a minimum in Problem 3.1 indeed exists. It is formulated as Proposition 3.2 below, whose proof, requiring results from Section 5, is given in Appendix D.

Proposition 3.2. *There exist matrices $H^* \in \mathbb{R}^{n\times k}$, and nonnegative diagonal $D^* \in \mathbb{R}^{n\times n}$, that minimize the I-divergence in Problem 3.1.*

In a statistical setup, the approximation problem has an equivalent formulation as an *estimation* problem. One then will have a sequence of *idd* observations $Y_1, \ldots, Y_N$, each distributed according to (2.7). The matrices H and D are the unknown parameters that have to be estimated, which can be done applying the maximum likelihood (ML) method. For big enough N, the sample covariance matrix will be positive definite a.s. under the assumption that the covariance matrix of the Y_i is positive definite. Denote the sample covariance matrix by $\widehat{\Sigma}$. The computation of the ML estimators of H and D is equivalent to solving the minimization problem 3.1. Indeed the normal log likelihood $\ell(H, D)$ with H and D as parameters yields

$$\ell(H, D) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|HH^\top + D| - \frac{1}{2}\mathrm{tr}\Big((HH^\top + D)^{-1}\widehat{\Sigma}\Big).$$

One immediately sees that $\ell(H, D)$ is, up to constants not depending on H and D, equal to $-\mathcal{I}(\widehat{\Sigma}||HH^\top + D)$. Hence, maximum likelihood estimation completely parallels I-divergence minimization, only the interpretation is different.

The equations for the maximum likelihood estimators can be found in e.g. Section 14.3.1 of Anderson (1984). In terms of the unknown parameters H and D, with D assumed to be non-singular, they are

$$H = (\widehat{\Sigma} - HH^\top)D^{-1}H \tag{3.4}$$

$$D = \Delta(\widehat{\Sigma} - HH^\top). \tag{3.5}$$

where $\Delta(M)$, defined for any square M, coincides with M on the diagonal and is zero elsewhere. Note that the matrix $HH^\top + D$ obtained by maximum likelihood estimation, is automatically invertible. Then it can be verified that equation (3.4) is equivalent to

$$H = \widehat{\Sigma}(HH^\top + D)^{-1}H, \tag{3.6}$$

which is also meaningful, when D is not invertible.

The maximum likelihood equations (3.4) and (3.5) for the alternative parametrization, as induced by (2.3), take the form

$$L = (\widehat{\Sigma} - LPL^\top)D^{-1}L \tag{3.7}$$

$$D = \Delta(\widehat{\Sigma} - LPL^\top), \tag{3.8}$$

with (3.7) equivalent to

$$L = \widehat{\Sigma}(LPL^\top + D)^{-1}L. \tag{3.9}$$

It is clear that the system of equations (3.4), (3.5) does not have an explicit solution. For this reason numerical algorithms have been devised, among others an adapted version of the EM algorithm, see Rubin and Thayer (1982). In the present paper we consider an alternative approach and, in Section 5, we compare the ensuing algorithm with the EM.

In Finesso and Spreij (2006) we considered an approximate nonnegative matrix factorization problem, where the objective function was also of I-divergence type. In that case, a relaxation technique lifted the original minimization to a double minimization in a higher dimensional space and led naturally to an alternating minimization algorithm. A similar approach, containing the core of the present paper, will be followed below.

4. Lifted version of the problem

In this section we recast Problem 3.1 in a higher dimensional space, making it amenable to solution via two partial minimizations. Later on this approach will lead to an alternating minimization algorithm.

First we introduce two relevant classes of normal distributions. All random vectors are supposed to be zero mean and normal, therefore their laws are completely specified by covariance matrices. Consider the set $\mathbf{\Sigma}$ comprising all the $(n+k)$-dimensional covariance matrices. An element $\Sigma \in \mathbf{\Sigma}$ can always be decomposed as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{4.1}$$

where Σ_{11} and Σ_{22} are square, of respective sizes n and k. Two subsets of $\mathbf{\Sigma}$ will play a major role in what follows. The subset $\mathbf{\Sigma}_0$ of $\mathbf{\Sigma}$, contains the covariances that can be written as in (4.1), with $\Sigma_{11} = \widehat{\Sigma}$, a given matrix, i.e.

$$\mathbf{\Sigma}_0 = \{\Sigma \in \mathbf{\Sigma} : \Sigma_{11} = \widehat{\Sigma}\}.$$

Elements of $\mathbf{\Sigma}_0$ will often be denoted by Σ_0. Also of interest is the subset $\mathbf{\Sigma}_1$ of $\mathbf{\Sigma}$ whose elements are covariances for which the decomposition (4.1) takes the special form

$$\Sigma = \begin{pmatrix} HH^\top + D & HQ \\ (HQ)^\top & Q^\top Q \end{pmatrix}, \tag{4.2}$$

for certain matrices H, D, Q with D diagonal, i.e.

$$\mathbf{\Sigma}_1 = \{\Sigma \in \mathbf{\Sigma} : \exists H, D, Q : \Sigma_{11} = HH^\top + D, \Sigma_{12} = HQ, \Sigma_{22} = Q^\top Q\}.$$

Elements of $\mathbf{\Sigma}_1$ will be often denoted by $\Sigma(H, D, Q)$ or by Σ_1.

In the present section we study the lifted

Problem 4.1.

$$\min_{\Sigma_0 \in \mathbf{\Sigma}_0, \Sigma_1 \in \mathbf{\Sigma}_1} \mathcal{I}(\Sigma_0 || \Sigma_1)$$

viewing it as a double minimization over the variables Σ_0 and Σ_1. Problem 4.1 and Problem 3.1 are related by the following proposition, whose proof is deferred to Appendix D.

Proposition 4.2. *Let $\widehat{\Sigma}$ be given. It holds that*

$$\min_{H,D} \mathcal{I}(\widehat{\Sigma} \,||\, HH^\top + D) = \min_{\Sigma_0 \in \mathbf{\Sigma}_0, \Sigma_1 \in \mathbf{\Sigma}_1} \mathcal{I}(\Sigma_0 || \Sigma_1).$$

4.1. Partial minimization problems

The first partial minimization, required for the solution of Problem 4.1, is as follows.

Problem 4.3. Given a strictly positive definite covariance matrix $\Sigma \in \mathbf{\Sigma}$, find

$$\min_{\Sigma_0 \in \mathbf{\Sigma}_0} \mathcal{I}(\Sigma_0 || \Sigma).$$

The unique solution to this problem can be computed analytically.

Proposition 4.4. *The unique minimizer Σ^* of Problem 4.3 is given by*

$$\Sigma^* = \begin{pmatrix} \widehat{\Sigma} & \widehat{\Sigma}\Sigma_{11}^{-1}\Sigma_{12} \\ \Sigma_{21}\Sigma_{11}^{-1}\widehat{\Sigma} & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}(\Sigma_{11} - \widehat{\Sigma})\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix} > 0.$$

Moreover

$$\mathcal{I}(\Sigma^* || \Sigma) = \mathcal{I}(\widehat{\Sigma} || \Sigma_{11}), \tag{4.3}$$

and the Pythagorean rule

$$\mathcal{I}(\Sigma_0 || \Sigma) = \mathcal{I}(\Sigma_0 || \Sigma^*) + \mathcal{I}(\Sigma^* || \Sigma) \tag{4.4}$$

holds for any strictly positive $\Sigma_0 \in \mathbf{\Sigma}_0$.

Proof. See Appendix D. □

Remark 4.5. Using the decomposition of Lemma B.1, one can easily compute the inverse of the matrix Σ^* of Proposition 4.4 and verify that $(\Sigma^*)^{-1}$ differs from Σ^{-1} only in the upper left block. Moreover, in terms of L^2-norms (the L^2-norm of a matrix M is $||M|| = (\mathrm{tr}(M^\top M))^{1/2}$) we have for the approximation of the inverse the identity $||\Sigma^{-1} - (\Sigma^*)^{-1}|| = ||\Sigma_{11}^{-1} - \widehat{\Sigma}^{-1}||$.

Next we turn to the second partial minimization

Problem 4.6. Given a strictly positive definite covariance matrix $\Sigma \in \mathbf{\Sigma}$, find

$$\min_{\Sigma_1 \in \mathbf{\Sigma}_1} \mathcal{I}(\Sigma||\Sigma_1).$$

A solution to this problem is given explicitly in the proposition below. To state the result we introduce the following notation: for any nonnegative definite P denote by $P^{1/2}$ any matrix satisfying ${P^{1/2}}^\top P^{1/2} = P$, and by $P^{-1/2}$ its inverse, if it exists. Furthermore we put $\tilde{\Sigma}_{11} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Proposition 4.7. *A minimizer $\Sigma(H^*, D^*, Q^*)$ of Problem 4.6 is given by*

$$\begin{aligned} Q^* &= \Sigma_{22}^{1/2} \\ H^* &= \Sigma_{12}\Sigma_{22}^{-1/2} \\ D^* &= \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}), \end{aligned}$$

corresponding to the minimizing matrix

$$\Sigma^* = \Sigma(H^*, D^*, Q^*) = \begin{pmatrix} \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Moreover, $\mathcal{I}(\Sigma||\Sigma^*) = \mathcal{I}(\tilde{\Sigma}_{11}||\Delta(\tilde{\Sigma}_{11}))$ and the Pythagorean rule

$$\mathcal{I}(\Sigma||\Sigma_1) = \mathcal{I}(\Sigma||\Sigma^*) + \mathcal{I}(\Sigma^*||\Sigma_1) \tag{4.5}$$

holds for any $\Sigma_1 = \Sigma(H, D, Q) \in \mathbf{\Sigma}_1$.

Proof. See Appendix D. □

Note that this problem cannot have a unique solution in terms of the matrices H and Q. Indeed, if U is a unitary $k \times k$ matrix and $H' = HU$, $Q' = U^\top Q$, then $H'H'^\top = HH^\top$, $Q'^\top Q' = Q^\top Q$ and $H'Q' = HQ$. Nevertheless, the optimal matrices $HH^\top$, HQ and $Q^\top Q$ are unique, as it can be easily checked using the expressions in Proposition 4.7.

Remark 4.8. Note that, since Σ is supposed to be strictly positive, $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} > 0$. It follows that $D^* = \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$ is strictly positive.

Remark 4.9. The matrix Σ^* in Proposition 4.7 differs from Σ only in the upper left block and in terms of L^2-norms we have the identity $||\Sigma - \Sigma^*|| = ||\tilde{\Sigma}_{11} - \Delta(\tilde{\Sigma}_{11})||$, compare with Remark 4.5.

We close this section by considering a constrained version of the second partial minimization Problem 4.6. The constraint that we impose is $Q = Q_0$, where Q_0 is fixed or, slightly more general, with $P_0 := Q_0^\top Q_0$ fixed. The matrices H and D remain free. For clarity we state this as

Problem 4.10. Given strictly positive covariances $\Sigma \in \mathbf{\Sigma}$ and $P_0 \in \mathbb{R}^{k\times k}$, and letting Q_0 be any matrix satisfying $P_0 = Q_0^\top Q_0$, find

$$\min_{\Sigma(H,D,Q_0)\in\mathbf{\Sigma}_1} \mathcal{I}(\Sigma||\Sigma_1).$$

The solution is given in the next proposition.

Proposition 4.11. *A solution Σ_0^* of Problem 4.10 is given by*

$$\Sigma_0^* = \begin{pmatrix} \Sigma_{12}\Sigma_{22}^{-1}P_0\Sigma_{22}^{-1}\Sigma_{21} + \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) & \Sigma_{12}\Sigma_{22}^{-1}P_0 \\ P_0\Sigma_{22}^{-1}\Sigma_{21} & P_0 \end{pmatrix},$$

for which $H^* = \Sigma_{12}\Sigma_{22}^{-1}Q_0^\top$ and D^* is as in Proposition 4.7.

Proof. See Appendix D. □

Note that for the constrained problem no Pythagorean rule holds. However (4.5) can be used to compare the optimal I-divergences of Problem 4.6 and Problem 4.10. Since $\Sigma_0^* \in \mathbf{\Sigma}_1$, applying (4.5) one gets

$$\mathcal{I}(\Sigma||\Sigma_0^*) = \mathcal{I}(\Sigma||\Sigma^*) + \mathcal{I}(\Sigma^*||\Sigma_0^*),$$

hence $\mathcal{I}(\Sigma||\Sigma_0^*) \geq \mathcal{I}(\Sigma||\Sigma^*)$, where Σ^* is as in Proposition 4.7. The quantity $\mathcal{I}(\Sigma^*||\Sigma_0^*)$ is the extra cost incurred solving Problem 4.10 instead of Problem 4.6. An elementary computation gives

$$\mathcal{I}(\Sigma^*||\Sigma_0^*) = \mathcal{I}(\Sigma_{22}||P_0).$$

In fact this is an easy consequence of the relation, similar to Remark 4.5,

$$(\Sigma_0^*)^{-1} - (\Sigma^*)^{-1} = \begin{pmatrix} 0 & 0 \\ 0 & P_0^{-1} - \Sigma_{22}^{-1} \end{pmatrix}.$$

We see that the two optimizing matrices in the constrained case (Proposition 4.11) and unconstrained case (Proposition 4.7) coincide iff the constraining matrix P_0 satisfies $P_0 = \Sigma_{22}$.

5. Alternating minimization algorithm

In this section, the core of the paper, the two partial minimizations of Section 4 are combined into an alternating minimization algorithm for the solution of Problem 3.1. A number of equivalent formulations of the updating equations will be presented and their properties discussed.

5.1. The algorithm

We suppose that the given covariance matrix $\widehat{\Sigma}$ is strictly positive definite. To setup the iterative minimization algorithm, assign initial values H_0, D_0, Q_0 to

the parameters, with D_0 diagonal, Q_0 invertible and $H_0H_0^\top + D_0$ invertible. The updating rules are constructed as follows. Let H_t, D_t, Q_t be the parameters at the t-th iteration, and $\Sigma_{1,t} = \Sigma(H_t, D_t, Q_t)$ the corresponding covariance, defined as in (4.2). Now solve the two partial minimizations as illustrated below.

$$(H_t, D_t, Q_t) \xrightarrow[\min_{\Sigma_0 \in \mathbf{\Sigma}_0} \mathcal{I}(\Sigma_0 || \Sigma_{1,t})]{\text{Prop. } 4.4} \Sigma_{0,t} \xrightarrow[\min_{\Sigma_1 \in \mathbf{\Sigma}_1} \mathcal{I}(\Sigma_{0,t} || \Sigma_1)]{\text{Prop. } 4.7} (H_{t+1}, D_{t+1}, Q_{t+1}) \cdots,$$

where $\Sigma_{0,t}$ denotes the solution of the first minimization with input $\Sigma_{1,t}$.
To express in a compact form the resulting update equations, define

$$R_t = I - H_t^\top (H_t H_t^\top + D_t)^{-1} H_t + H_t^\top (H_t H_t^\top + D_t)^{-1} \widehat{\Sigma} (H_t H_t^\top + D_t)^{-1} H_t. \quad (5.1)$$

Note that, by Remark 4.8, $H_tH_t^\top + D_t$ is actually invertible for all t, since both $H_0H_0^\top + D_0$ and Q_0 have been chosen to be invertible. It follows, by Corollary B.4, that also $I - H_t^\top (H_tH_t^\top + D_t)^{-1}H_t$, and consequently R_t, are strictly positive and therefore invertible. The update equations resulting from the cascade of the two minimizations are

$$Q_{t+1} = \left(Q_t^\top R_t Q_t\right)^{1/2}, \quad (5.2)$$

$$H_{t+1} = \widehat{\Sigma}(H_tH_t^\top + D_t)^{-1} H_t Q_t Q_{t+1}^{-1}, \quad (5.3)$$

$$D_{t+1} = \Delta(\widehat{\Sigma} - H_{t+1}H_{t+1}^\top). \quad (5.4)$$

Properly choosing the square root in Equation (5.2) makes Q_t disappear from the update equations. This is an attractive feature since only H_t and D_t are needed to construct the approximate FA model $H_tH_t^\top + D_t$ at the t-th step of the algorithm. Observe that $(Q_t^\top R_t Q_t)^{1/2} = R_t^{1/2} Q_t$, where $R_t^{1/2}$ is a symmetric root of R_t, is a possible root for the right hand side of Equation (5.2). Inserting the resulting matrix $Q_{t+1} = R_t^{1/2} Q_t$ into Equation (5.3) results in

Algorithm 5.1. Given H_t, D_t from the t-th step, and R_t as in (5.1), the update equations for a I-divergence minimizing algorithm are

$$H_{t+1} = \widehat{\Sigma}(H_tH_t^\top + D_t)^{-1} H_t R_t^{-1/2} \quad (5.5)$$

$$D_{t+1} = \Delta(\widehat{\Sigma} - H_{t+1}H_{t+1}^\top). \quad (5.6)$$

Since R_t only depends on H_t and D_t, see (5.1), the parameter Q_t has been effectively eliminated.

5.2. Alternative algorithms

Algorithm 5.1 has two drawbacks making its implementation computationally awkward. To update H_t via equation (5.5) one has to compute, at each step, the square root of the $k \times k$ matrix R_t and the inverse of the $n \times n$ matrix

$H_tH_t^\top + D_t$. Taking a slightly different approach it is possible to reorganize the algorithm in order to avoid the computation of square roots at each step, and to reduce to $k \times k$ the size of the matrices that need to be inverted.

To avoid the computation of square roots at each step there are at least two possible variants of Algorithm 5.1, both involving a reparametrization. The first approach is to use the alternative parametrization (2.3) and to write update equations for the parameters L, D, P. Translated in terms of the matrices $L_t := H_tQ_t^{-\top}$ and $P_t = Q_t^\top Q_t$, Algorithm 5.1 becomes

Algorithm 5.2. Given L_t, P_t, and D_t from the t-th step, the update equations for a I-divergence minimizing algorithm are

$$\begin{aligned} L_{t+1} &= \widehat{\Sigma}(L_tP_tL_t^\top + D_t)^{-1}L_tP_tP_{t+1}^{-1}, \\ P_{t+1} &= P_t - P_tL_t^\top(L_tP_tL_t^\top + D_t)^{-1}(L_tP_tL_t^\top + D_t - \widehat{\Sigma})(L_tP_tL_t^\top + D_t)^{-1}L_tP_t, \\ D_{t+1} &= \Delta(\widehat{\Sigma} - L_{t+1}P_{t+1}L_{t+1}^\top). \end{aligned} \tag{5.7}$$

One can run Algorithm 5.2 for any number T of steps, and then switch back to the H, D parametrization computing $H_T = L_TQ_T^\top$, which requires only the square root at iteration T, *i.e.* $P_T = Q_T^\top Q_T$

An alternative approach to avoid the square roots at each iteration of Algorithm 5.1 is to run it for $\mathcal{H}_t := H_tH_t^\top$.

Proposition 5.3. *Let H_t be as in Algorithm 5.1. Pick $\mathcal{H}_0 = H_0H_0^\top$, and D_0 such that $\mathcal{H}_0 + D_0$ is invertible. The update equation for $\mathcal{H}_t$ becomes*

$$\mathcal{H}_{t+1} = \widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t\big(D_t + \widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t\big)^{-1}\widehat{\Sigma}. \tag{5.8}$$

Proof. From Equation (5.5) one immediately gets

$$\mathcal{H}_{t+1} = H_{t+1}H_{t+1}^\top = \widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}H_tR_t^{-1}H_t^\top(\mathcal{H}_t + D_t)^{-1}\widehat{\Sigma}. \tag{5.9}$$

The key step in the proof is an application of the elementary identity

$$(I + H^\top PH)^{-1}H^\top = H^\top(I + PHH^\top)^{-1},$$

valid for all H and P of appropriate dimensions for which both inverses exist. Note that, by Corollary B.3, the two inverses either both exist or both do not exist. We have already seen that R_t is invertible and of the type $I + HPH^\top$. Following this recipe, we compute

$$\begin{aligned} R_t^{-1}H_t^\top &= H_t^\top\big(I - (\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t + (\mathcal{H}_t + D_t)^{-1}\widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t\big)^{-1} \\ &= H_t^\top\big((\mathcal{H}_t + D_t)^{-1}D_t + (\mathcal{H}_t + D_t)^{-1}\widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t\big)^{-1} \\ &= H_t^\top\big(D_t + \widehat{\Sigma}(\mathcal{H}_t + D_t)^{-1}\mathcal{H}_t\big)^{-1}(\mathcal{H}_t + D_t). \end{aligned}$$

Insertion of this result into (5.9) yields (5.8). □

One can run the update Equation (5.9), for any number T of steps, and then switch back to H_T, taking any $n \times k$ factor of $\mathcal{H}_T$ *i.e.* solve $\mathcal{H}_T = H_T H_T^\top$. Since Equation (5.9) transforms $\mathcal{H}_t$ into $\mathcal{H}_{t+1}$ preserving the rank, the latter factorization is always possible.

It is apparent that the second computational issue we mentioned above, concerning the inversion of $n \times n$ matrices at each step, affects also Algorithm 5.2. The alternative form of the update equations derived below only requires the inversion of $k \times k$ matrices: a very desirable property since k is usually much smaller than n. Referring to Algorithm 5.1, since D_t is invertible, apply Corollary B.2 to find

$$(H_t H_t^\top + D_t)^{-1} H_t = D_t^{-1} H_t (I + H_t^\top D_t^{-1} H_t)^{-1}.$$

The alternative expression for R_t is

$$R_t = (I + H_t^\top D_t^{-1} H_t)^{-1} + (I + H_t^\top D_t^{-1} H_t)^{-1} H_t^\top D_t^{-1} \widehat{\Sigma} D_t^{-1} H_t (I + H_t^\top D_t^{-1})^{-1}.$$

The update formula (5.5) can therefore be replaced with

$$H_{t+1} = \widehat{\Sigma} D_t^{-1} H_t (I + H_t^\top D_t^{-1} H_t)^{-1} R_t^{-1/2}.$$

Similar results can be derived also for Algorithm 5.2.

5.3. Asymptotic properties

In the portmanteau proposition below we collect the asymptotic properties of Algorithm 5.1, also quantifying the I-divergence decrease at each step.

Proposition 5.4. *For Algorithm 5.1 the following hold.*

(a) $H_t H_t^\top \leq \widehat{\Sigma}$ *for all* $t \geq 1$.
(b) If $D_0 > 0$ *and* $\Delta(\widehat{\Sigma} - D_0) > 0$ *then* $D_t > 0$ *for all* $t \geq 1$.
(c) The matrices R_t *are invertible for all* $t \geq 1$.
(d) If $H_t H_t^\top + D_t = \widehat{\Sigma}$ *then* $H_{t+1} = H_t$, $D_{t+1} = D_t$.
(e) Decrease of the objective function:

$$\mathcal{I}(\widehat{\Sigma} || \widehat{\Sigma}_t) - \mathcal{I}(\widehat{\Sigma} || \widehat{\Sigma}_{t+1}) = \mathcal{I}(\Sigma_{1,t+1} || \Sigma_{1,t}) + \mathcal{I}(\Sigma_{0,t} || \Sigma_{0,t+1}),$$

where $\widehat{\Sigma}_t = H_t H_t^\top + D_t$ *is the* t*-th approximation of* $\widehat{\Sigma}$*, and* $\Sigma_{0,t}, \Sigma_{1,t}$ *were defined in subsection 5.1.*
(f) The interior limit points (H, D) *of the algorithm satisfy*

$$H = (\widehat{\Sigma} - H H^\top) D^{-1} H, \qquad D = \Delta(\widehat{\Sigma} - H H^\top), \tag{5.10}$$

which are the ML equations (3.4) and (3.5). If (H, D) *is a solution to these equation also* (HU, D) *is a solution, for any unitary matrix* $U \in \mathbb{R}^{k \times k}$.
(g) Limit points $(\mathcal{H}, D)$*, see (5.9), satisfy*

$$\mathcal{H} = \widehat{\Sigma}(\mathcal{H} + D)^{-1} \mathcal{H}, \qquad D = \Delta(\widehat{\Sigma} - \mathcal{H}).$$

Proof. (a) This follows from Remark 4.8 and the construction of the algorithm as a combination of the two partial minimizations.
(b) This similarly follows from Remark 4.8.
(c) Use the identity $I - H_t^\top (H_t H_t^\top + D_t)^{-1} H_t = (I + H_t^\top D_t^{-1} H_t)^{-1}$ and $\widehat{\Sigma}$ nonnegative definite.
(d) In this case, Equation (5.1) shows that $R_t = I$ and substituting this into the update equations yields the conclusion.
(e) As matter of fact, we can express the decrease as a sum of two I-divergences, since the algorithm is the superposition of the two partial minimization problems. The results follows from a concatenation of Proposition 4.4 and Proposition 4.7.
(f) We consider Algorithm 5.2 first. Assume that all variables converge. Then, from (5.7), for limit points L, P, D it holds that

$$L = \widehat{\Sigma}(LPL^\top + D)^{-1} L,$$

which coincides with equation (3.4). Let then Q be a square root of P and $H = LQ^\top$. This gives the first of the desired relations. The rest is trivial.
(g) This follows by inserting the result of (f). □

In part (f) of Proposition 5.4 we have made the assumption that the limit points are interior points. This assumption does not always hold true, it may happen that a limit point (H, D) is such that D contains zeros on the diagonal. We will treat this extensively in Section 7.1 in connection with a restricted optimization problem, in which it is imposed that D has a number of zeros on the diagonal.

6. Comparison with the EM algorithm

Rubin and Thayer (1982) put forward a version of the EM algorithm (see Dempster, Laird and Rubin (1977)) in the context of estimation for FA models. Their algorithm is as follows.

Algorithm 6.1 (EM)**.**

$$H_{t+1} = \widehat{\Sigma}(H_t H_t^\top + D_t)^{-1} H_t R_t^{-1} \tag{6.1}$$

$$D_{t+1} = \Delta(\widehat{\Sigma} - H_{t+1} R_t H_{t+1}^\top), \tag{6.2}$$

where $R_t = I - H_t^\top (H_t H_t^\top + D_t)^{-1} (H_t H_t^\top + D_t - \widehat{\Sigma})(H_t H_t^\top + D_t)^{-1} H_t$.

The EM Algorithm 6.1 differs in both equations from our Algorithm 5.1. It is well known that EM algorithms can be derived as alternating minimizations, see Csiszár and Tusnády (1984), it is therefore interesting to investigate how Algorithm 6.1 can be derived within our framework. Thereto one considers the first partial minimization problem together with the *constrained* second partial minimization Problem 4.10, the constraint being $Q = Q_0$, for some Q_0. Later on we will see that the particular choice of Q_0, as long as it is invertible, is irrelevant.

The concatenation of these two problems results in the EM Algorithm 6.1, as is detailed below.

Starting at (H_t, D_t, Q_0), one performs the first partial minimization, that results in the matrix

$$\begin{pmatrix} \widehat{\Sigma} & \widehat{\Sigma}(H_tH_t + D_t)^{-1}H_tQ_0 \\ Q_0^\top H_t^\top (H_tH_t + D_t)^{-1}\widehat{\Sigma} & Q_0^\top R_tQ_0 \end{pmatrix}.$$

Performing now the *constrained* second minimization, according to the results of Proposition 4.11, one obtains

$$H_{t+1} = \widehat{\Sigma}(H_tH_t^\top + D_t)^{-1}H_tR_t^{-1} \tag{6.3}$$

$$D_{t+1} = \Delta\big(\widehat{\Sigma} - \widehat{\Sigma}(H_tH_t^\top + D_t)^{-1}H_tR_t^{-1}H_t^\top(H_tH_t^\top + D_t)^{-1}\widehat{\Sigma}\big). \tag{6.4}$$

Substitution of (6.3) into (6.4) yields

$$D_{t+1} = \Delta(\widehat{\Sigma} - H_{t+1}R_tH_{t+1}^\top).$$

One sees that the matrix Q_0 does not appear in the recursion, just as the matrices Q_t do not occur in Algorithm 5.1.

Both Algorithms 5.1 and 6.1 are the result of two partial minimization problems. The latter algorithm differs from ours in that the second partial minimization is *constrained*. It is therefore reasonable to expect that, from the point of view of minimizing I-divergence, Algorithm 5.1 yields a better performance, although comparisons must take into account that the initial parameters for the two *species* of the second partial minimization will in general be different. We will illustrate these considerations by some numerical examples in Section 8.

We also note that for Algorithm 5.1 it was possible to identify the update gain at each step, see Proposition 5.4(e), resulting from the two Pythagorean rules. For the EM algorithm a similar formula cannot be given, because for the constrained second partial minimization a Pythagorean rule does not hold, see the discussion after Proposition 4.11 in Section 4.1.

7. Singular D

It has been known for a long time, see e.g. Jöreskog (1967), that numerical solutions to the ML equations (see Section 3) often produce a nearly singular matrix D. This motivates the investigation of the stationary points (H, D) of Algorithm 5.1 with singular D, *i.e.* with zeros on the diagonal (Section 7.1). Naturally connected to this is the analysis of the minimization Problem 3.1 when D is *constrained*, at the outset, to be singular (Section 7.2), and the investigation of its consequences for the minimization algorithm of Proposition 5.3 (Section 7.3).

7.1. Stationary points (H, D) with singular D

As mentioned before, already in Jöreskog (1967) it has been observed that, numerically maximizing the likelihood, one often reaches matrices D that are nearly singular. This motivates the investigation of the stationary points (H, D) of Algorithm 5.1 for which D is singular, i.e.

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} = \begin{pmatrix} D_1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{7.1}$$

where $D_1 > 0$ has size $n_1 \times n_1$ and the lower right zero block has size $n_2 \times n_2$, with $n_1 + n_2 = n$.
Accordingly we partition $H \in \mathbb{R}^{n\times k}$ as

$$H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix}, \tag{7.2}$$

where $H_1 \in \mathbb{R}^{n_1\times k}$ and $H_2 \in \mathbb{R}^{n_2\times k}$. Then

$$HH^\top + D = \begin{pmatrix} H_1H_1^\top + D_1 & H_1H_2^\top \\ H_2H_1^\top & H_2H_2^\top \end{pmatrix}. \tag{7.3}$$

We recall that Problem 3.1 calls for the minimization, over H and D, of the functional $\mathcal{I}(\widehat{\Sigma}||HH^\top + D)$, which is finite if and only if $HH^\top + D$ is strictly positive definite. In view of (7.3), this happens if and only if

$$H_2H_2^\top > 0,$$

the standing assumption of this section. A direct consequence of this assumption is that $n_2 \leq k$.
The *given* matrix $\widehat{\Sigma}$ will be similarly decomposed as

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22}. \end{pmatrix} \tag{7.4}$$

Proposition 7.1. *If (H, D) is a stationary point of the algorithm, with D as in (7.1), then the given matrix $\widehat{\Sigma}$ is such that $\widehat{\Sigma}_{22} = H_2H_2^\top$ and $\widehat{\Sigma}_{12} = H_1H_2^\top$.*

Proof. By Proposition 5.4 $\widehat{\Sigma} - HH^\top$ is nonnegative definite, as is its lower right block $\widehat{\Sigma}_{22} - H_2H_2^\top$. Since $D = \Delta(\widehat{\Sigma} - HH^\top)$ and $D_2 = 0$, we get that $\Delta(\widehat{\Sigma}_{22} - H_2H_2^\top) = 0$ and therefore $\widehat{\Sigma}_{22} = H_2H_2^\top$. We conclude that

$$\widehat{\Sigma} - HH^\top = \begin{pmatrix} \widehat{\Sigma}_{11} - H_1H_1^\top & \widehat{\Sigma}_{12} - H_1H_2^\top \\ \widehat{\Sigma}_{21} - H_2H_1^\top & 0 \end{pmatrix} \geq 0,$$

hence $\widehat{\Sigma}_{12} = H_1H_2^\top$. □

Define

$$\widetilde{H}_1 := H_1(I - H_2^\top (H_2 H_2^\top)^{-1} H_2). \tag{7.5}$$

Since $I - H_2^\top (H_2 H_2^\top)^{-1} H_2$ is a projection, one finds

$$\widetilde{H}_1 \widetilde{H}_1^\top = H_1(I - H_2^\top (H_2 H_2^\top)^{-1} H_2) H_1^\top. \tag{7.6}$$

In view of Proposition 7.1 this becomes

$$\widetilde{H}_1 \widetilde{H}_1^\top = H_1 H_1^\top - \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21}. \tag{7.7}$$

Finally we need

$$\widetilde{\Sigma}_{11} := \widehat{\Sigma}_{11} - \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21}. \tag{7.8}$$

Proposition 7.2. *If (H, D) is a stationary point of the algorithm with $D_2 = 0$, then*

$$\mathcal{I}(\widehat{\Sigma} || HH^\top + D) = \mathcal{I}(\widetilde{\Sigma}_{11} || \widetilde{H}_1 \widetilde{H}_1^\top + D_1).$$

Moreover, the stationary equations (5.10) reduce to

$$\begin{aligned} \widetilde{H}_1 &= \widetilde{\Sigma}_{11} (\widetilde{H}_1 \widetilde{H}_1^\top + D_1)^{-1} \widetilde{H}_1 = (\widetilde{\Sigma}_{11} - \widetilde{H}_1 \widetilde{H}_1^\top) D_1^{-1} \widetilde{H}_1 \\ D_1 &= \Delta(\widetilde{\Sigma}_{11} - \widetilde{H}_1 \widetilde{H}_1^\top). \end{aligned}$$

Proof. One easily verifies that for any nonsingular matrix A of the appropriate size

$$\mathcal{I}(APA^\top || AQA^\top) = \mathcal{I}(P||Q).$$

Taking

$$A = \begin{pmatrix} I & -\widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \\ 0 & I \end{pmatrix},$$

one finds

$$A \widehat{\Sigma} A^\top = \begin{pmatrix} \widetilde{\Sigma}_{11} & 0 \\ 0 & \widehat{\Sigma}_{22} \end{pmatrix}.$$

Moreover, by Proposition 7.1,

$$A(HH^\top + D)A^\top = \begin{pmatrix} H_1 H_1 + D_1 - \widehat{\Sigma}_{12} \widehat{\Sigma}_{22}^{-1} \widehat{\Sigma}_{21} & 0 \\ 0 & \widehat{\Sigma}_{22}, \end{pmatrix}.$$

where the upper left block is equal to $\widetilde{H}_1 \widetilde{H}_1^\top + D_1$ in view of equation (7.7). The first assertion follows. The reduced stationary equations follow by simple computation. □

Remark 7.3. Under the conditions of Proposition 7.2, the pair $(\widetilde{H}_1, D_1)$ is also a stationary point for the minimization of $\mathcal{I}(\widetilde{\Sigma}_{11} || \widetilde{H}_1 \widetilde{H}_1^\top + D_1)$. This is in full agreement with the results of Section 7.2.

7.2. Approximation with singular D

In this section we consider the approximation Problem 3.1 under the constraint $D_2 = 0$. Jöreskog (1967) investigated the solution of the likelihood equations (3.5) and (3.6) under zero constraints on D, whereas in this section we work directly on the objective function of Problem 3.1 without referring to those equations. The constrained minimization problem can be formulated as

Problem 7.4. Given $\widehat{\Sigma} > 0$ of size $n \times n$ and integers n_2 and k, with $n_2 \leq k < n$, minimize

$$\mathcal{I}(\widehat{\Sigma} || HH^\top + D), \tag{7.9}$$

over (H, D) with D satisfying (7.1).

We will now decompose the objective function, choosing a convenient representation of the matrix H, in order to reduce the complexity of Problem 7.4. To that end we make the following observation. Given any orthogonal matrix Q, define $H' = HQ$, then clearly $H'H'^\top + D = HH^\top + D$. Let $H_2 = U(0\ \Lambda)V^\top$ be the singular value decomposition of H_2, with Λ a positive definite diagonal matrix of size $n_2 \times n_2$, and U and V orthogonal of sizes $n_2 \times n_2$ and $k \times k$ respectively. Let

$$H' = HV$$

The blocks of H' are $H'_1 = H_1 V$ and $H'_2 = (H'_{21}\, H'_{22}) := (0 \quad U\Lambda)$, with $H'_{21} \in \mathbb{R}^{(k-n_2)\times n_2}$ and $H'_{22} \in \mathbb{R}^{n_2 \times n_2}$. Hence, without loss of generality, in the remainder of this section we assume that

$$H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix}, \qquad H_{22} \text{ invertible.} \tag{7.10}$$

Finally, let

$$K = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1} - H_1 H_2^\top (H_2 H_2^\top)^{-1},$$

which, under (7.10), is equivalent to

$$K = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1} - H_{12}H_{22}^{-1}.$$

Here is the announced decomposition of the objective function.

Lemma 7.5. *Let D be as in equation (7.1). The following I-divergence decomposition holds.*

$$\begin{aligned} \mathcal{I}(\widehat{\Sigma} || HH^\top + D) = {} & \mathcal{I}(\widetilde{\Sigma}_{11} || H_{11}H_{11}^\top + D_1) + \mathcal{I}(\widehat{\Sigma}_{22} || H_{22}H_{22}^\top) \\ & + \frac{1}{2}\mathrm{tr}\big(\widehat{\Sigma}_{22}K^\top (H_{11}H_{11}^\top + D_1)^{-1}K\big). \end{aligned} \tag{7.11}$$

Proof. See Appendix D. □

We are now ready to characterize the solution of Problem 7.4.

Proposition 7.6. *Any pair (H, D), as in (7.1) and (7.10), solving Problem 7.4 satisfies*

- $\mathcal{I}(\widetilde{\Sigma}_{11}||H_{11}H_{11}^\top + D_1)$ *is minimized,*
- $H_{22}H_{22}^\top = \widehat{\Sigma}_{22}$,
- $H_{12} = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}H_{22}$.

Proof. Observe first that the second and third terms on the right hand side of (7.11) are nonnegative and can be made zero. To this end it is enough to select H_{22} such that $H_{22}H_{22}^\top = \widehat{\Sigma}_{22}$ and then $H_{12} = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}H_{22}$. The remaining blocks, H_{11} and D_1, are determined minimizing the first term. □

Remark 7.7. In the special case $n_2 = k$, the matrices H_{11} and H_{21} are empty, $H_{12} = H_1$, and $H_{22} = H_2$. From Proposition 7.6, at the minimum, $H_2H_2^\top = \widehat{\Sigma}_{22}$, $H_1H_2^\top = \widehat{\Sigma}_{12}$, and D_1 minimizes $\mathcal{I}(\widetilde{\Sigma}_{11}||D_1)$. The latter problem has solution $D_1 = \Delta(\widetilde{\Sigma}_{11})$. It is remarkable that in this case the minimization problem has an *explicit* solution.

Proposition 7.6 also sheds some light on the unconstrained Problem 3.1.

Corollary 7.8. *Assume that, in Problem 3.1, $\mathcal{I}(\widehat{\Sigma}||HH^\top + D)$ is minimized for a pair (H, D) with D of the form (7.1). Then $\widehat{\Sigma}_{12} = H_1H_2^\top$, $\widehat{\Sigma}_{22} = H_2H_2^\top$, and $(\widetilde{H}_1, D_1)$, where $\widetilde{H}_1$ is as in (7.5), minimizes $\mathcal{I}(\widetilde{\Sigma}_{11}||\widetilde{H}_1\widetilde{H}_1^\top + D_1)$.*

Proof. It is obvious that, in this case, Problem 3.1 and Problem 7.4 are equivalent. The result readily follows from Proposition 7.6, in view of the equality $\widetilde{H}_1 = (H_{11}\, 0)$. □

Hence, in Problem 3.1, a singular minimizer D can occur only if $\widehat{\Sigma}$ has the special structure described in Corollary 7.8.

In the same spirit one can characterize the covariances $\widehat{\Sigma}$, admitting an exact FA model of size $k \geq n_2$, and with $D_2 = 0$. This happens if and only if, with the notations of Section 7.1,

$$\widehat{\Sigma}_{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21} \quad \text{is a diagonal matrix.} \tag{7.12}$$

This condition is easily interpreted in terms of random vectors. Let Y be an n dimensional, zero mean, normal vector with $\mathbb{C}\mathrm{ov}(Y) = \widehat{\Sigma}$ and partition it into two subvectors (Y_1, Y_2), of respective sizes n_1 and n_2, corresponding to the block partitioning of $\widehat{\Sigma}$. The above condition states that the components of Y_1 are conditionally independent given Y_2. The construction of the k-dimensional, exact FA model, with $D_2 = 0$ is as follows.

Let $D_1 = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, which is a diagonal matrix by assumption. Let R be the symmetric, invertible, square root of Σ_{22}. Define the matrices

$$H_1 = \Sigma_{12}\begin{pmatrix} R^{-1} & 0 \end{pmatrix} \in \mathbb{R}^{n_1 \times k}$$
$$H_2 = \begin{pmatrix} R & 0 \end{pmatrix} \in \mathbb{R}^{n_2 \times k}.$$

One verifies the identities $H_2H_2^\top = \Sigma_{22}$, $H_1H_2^\top = \Sigma_{12}$ and $H_1H_1^\top = \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. It follows that $\Sigma_{11} = D_1 + H_1H_1^\top$. Let Z be a $(k-n_2)$-dimensional random vector, independent of Y, with zero mean and identity covariance matrix. Put

$$X = \begin{pmatrix} R^{-1}Y_2 \\ Z \end{pmatrix}.$$

Then $\mathbb{C}\mathrm{ov}(X) = I_k$. Furthermore, $\varepsilon_1 := Y_1 - H_1X$ is independent of X with $\mathbb{C}\mathrm{ov}(\varepsilon_1) = D_1$, and $Y_2 - H_2X = 0$. It follows that

$$\begin{aligned} Y_1 &= H_1X + \varepsilon_1 \\ Y_2 &= H_2X \end{aligned}$$

is an exact realization of Y in terms of a factor model.

7.3. Algorithm when a part of D has zero diagonal

In Section 7.2 we have posed the minimization problem under the additional constraint that the matrix D contains a number of zeros on the diagonal. In the present section we investigate how this constraint affects the alternating minimization algorithm. For simplicity we give a detailed account of this, only using the recursion (5.8) for $\mathcal{H}_t$. Initialize the algorithm at (H_0, D_0) with

$$D_0 = \begin{pmatrix} \widetilde{D} & 0 \\ 0 & 0 \end{pmatrix}, \tag{7.13}$$

where $\widetilde{D} > 0$ is of size $n_1 \times n_1$ and

$$H_0 = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix}, \tag{7.14}$$

where $H_2 \in \mathbb{R}^{n_2 \times k}$ is assumed to have full row rank, so that $n_2 \leq k$ (note the slight ambiguity in the notation for the blocks of H_0). Clearly $H_0H_0^\top + D_0$ is invertible. For H_0 as in equation (7.14) put

$$\widetilde{\mathcal{H}} = H_1(I - H_2^\top(H_2H_2^\top)^{-1}H_2)H_1^\top. \tag{7.15}$$

We have the following result.

Lemma 7.9. *Let (H_0, D_0) be given as above, and $\mathcal{H}_0 = H_0H_0^\top$. Applying one step of recursion (5.8), one gets*

$$\mathcal{H}_1 = \begin{pmatrix} \mathcal{H}^{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{pmatrix}, \tag{7.16}$$

where

$$\mathcal{H}^{11} = \widetilde{\Sigma}_{11}(\widetilde{\mathcal{H}} + \widetilde{D})^{-1}\widetilde{\mathcal{H}}(\widetilde{D} + \widetilde{\Sigma}_{11}(\widetilde{\mathcal{H}} + \widetilde{D})^{-1}\widetilde{\mathcal{H}})^{-1}\widetilde{\Sigma}_{11} + \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}. \tag{7.17}$$

and

$$D_1 = \begin{pmatrix} \Delta(\widetilde{\Sigma}_{11} - \widetilde{\mathcal{H}}) & 0 \\ 0 & 0 \end{pmatrix}. \tag{7.18}$$

Proof. We start from Equation (5.8) with $t = 0$ and compute the value of $\mathcal{H}_1$. To that end we first obtain under the present assumption an expression for the matrix $(\mathcal{H} + D_0)^{-1}\mathcal{H}$. Let $P = I - H_2^\top (H_2 H_2^\top)^{-1} H_2$. It holds that

$$(\mathcal{H} + D_0)^{-1}\mathcal{H} = \begin{pmatrix} (\widetilde{D} + H_1 P H_1^\top)^{-1} H_1 P H_1^\top & 0 \\ (H_2 H_2^\top)^{-1} H_2 H_1^\top (\widetilde{D} + H_1 P H_1^\top)^{-1} \widetilde{D} & I \end{pmatrix}, \tag{7.19}$$

as one can easily verify by multiplying this equation by $\mathcal{H} + D_0$. We also need the inverse of $D_0 + \widehat{\Sigma}(\mathcal{H} + D_0)^{-1}\mathcal{H}$, postmultiplied with $\widehat{\Sigma}$. Introduce $U = \widetilde{D} + \widetilde{\Sigma}_{11}(H_1 P H_1^\top + \widetilde{D})^{-1} H_1 P H_1^\top$ and

$$V = \widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}(H_1 P H_1^\top + \widetilde{D})^{-1} + (H_2 H_2^\top)^{-1} H_2 H_1^\top (H_2 H_2^\top)^{-1}\widetilde{D}.$$

It results that

$$\big(D_0 + \widehat{\Sigma}(\mathcal{H} + D_0)^{-1}\mathcal{H}\big)^{-1}\widehat{\Sigma} = \begin{pmatrix} U^{-1}\widetilde{\Sigma}_{11} & 0 \\ -VU^{-1}\widetilde{\Sigma}_{11} + \widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21} & I \end{pmatrix}. \tag{7.20}$$

Insertion of the expressions (7.19) and (7.20) into (5.8) yields the result. □

The update equations of the algorithm for $\mathcal{H}_t$ and D_t, can be readily derived from Lemma 7.9 and are summarized below.

Proposition 7.10. *The upper left block $\mathcal{H}_t^{11}$ of $\mathcal{H}_t$, can be computed running a recursion for $\widetilde{\mathcal{H}}_t := \mathcal{H}_t^{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}$,*

$$\widetilde{\mathcal{H}}_{t+1} = \widetilde{\Sigma}_{11}(\widetilde{\mathcal{H}}_t + \widetilde{D}_t)^{-1}\widetilde{\mathcal{H}}_t(\widetilde{D}_t + \widetilde{\Sigma}_{11}(\widetilde{\mathcal{H}}_t + \widetilde{D}_t)^{-1}\widetilde{\mathcal{H}}_t)^{-1}\widetilde{\Sigma}_{11},$$

whereas the blocks on the border of $\mathcal{H}_t$ remain constant. The iterates for D_t all have a lower right block of zeros, while the upper left $n_1 \times n_1$ block $\widetilde{D}_t$ satisfies

$$\widetilde{D}_t = \Delta(\widetilde{\Sigma}_{11} - \widetilde{\mathcal{H}}_t).$$

□

Note that the recursions of Proposition 7.10 are exactly those that follow from the optimization Problem 7.4. Comparison with (5.8), shows that, while the algorithm for the unconstrained case updates $\mathcal{H}_t$ of size $n \times n$, now one needs to update $\widetilde{\mathcal{H}}_t$ which is of smaller size $n_1 \times n_1$.

Now we specialize the above to the case in which $n_2 = k$.

Corollary 7.11. *Let the initial value D_0 be as in Equation (7.13) with $n_2 = k$. Then for any initial value $\mathcal{H}_0$ the algorithm converges* in one step *and one has that the first iterates D_1 and $\mathcal{H}_1$, which are equal to the terminal values, are given by*

$$D_1 = \begin{pmatrix} \Delta(\widetilde{\Sigma}_{11}) & 0 \\ 0 & 0 \end{pmatrix}$$
$$\mathcal{H}_1 = \begin{pmatrix} \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{pmatrix}.$$

Proof. We use Proposition 7.9 and notice that in the present case the matrix $\widetilde{\mathcal{H}}$ of (7.15) is equal to zero. Therefore $\widetilde{\mathcal{H}}^{11} = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}$ and the result follows. □

It is remarkable that in this case we have convergence of the iterates in one step only. Moreover the resulting values are exactly the theoretical ones, which we have explicitly computed in Remark 7.7.

8. Numerical examples

8.1. Simulated data

In the present section we investigate the performance of the Algorithm 5.1 and compare it to the behaviour of the EM Algorithm 6.1. In all examples we take $\widehat{\Sigma}$ equal to $AA^\top + c\,\mathrm{diag}(d)$, where $A \in \mathbb{R}^{n\times m}$ with $m \leq n$, $d \in \mathbb{R}^n_+$ and $c \geq 0$, for various values of n, m. The matrices A and the vector d have been randomly generated. The notation A=rand(n, m) means that A is a randomly generated matrix of size $n \times m$, whose elements are independently drawn from the uniform distribution on $[0, 1]$. In all cases the resulting matrix $\widehat{\Sigma}$ is strictly positive definite. The reason for incorporating the component d is that we want to check whether the algorithm is able to reconstruct $\widehat{\Sigma} = AA^\top + \mathrm{diag}(d)$ in case the inner size k of the matrices H_t produced by the algorithm is taken to be equal to m.

We have also included results on the L^2-norm of the difference between the given matrix $\widehat{\Sigma}$ and its approximants $\Sigma_t = H_tH_t^\top + D_t$, i.e. we also compute $\ell_t = \big(\mathrm{tr}((\widehat{\Sigma} - \Sigma_t)^\top(\widehat{\Sigma} - \Sigma_t))\big)^{1/2}$. The origin of this extra means of comparison of behavior of the Algorithms 5.1 and 6.1 is that we detected in a number of cases that in terms of the value of the divergences, the difference between the approximations generated by the two algorithms was, after enough iterations, negligible, whereas a simple look at the matrices produced by the final iterations revealed that Algorithm 5.1 produced very acceptable, if not outstanding results, whereas the approximations generated by the EM algorithm 6.1 for the same given matrix were rather poor. This phenomenon is reflected by a huge L^2-error of the EM algorithm, as compared to a small one of Algorithm 5.1. The choice for the L^2-norm is to some extent arbitrary. We are basically concerned with good approximations in terms of I-divergence, and it is therefore a priori not

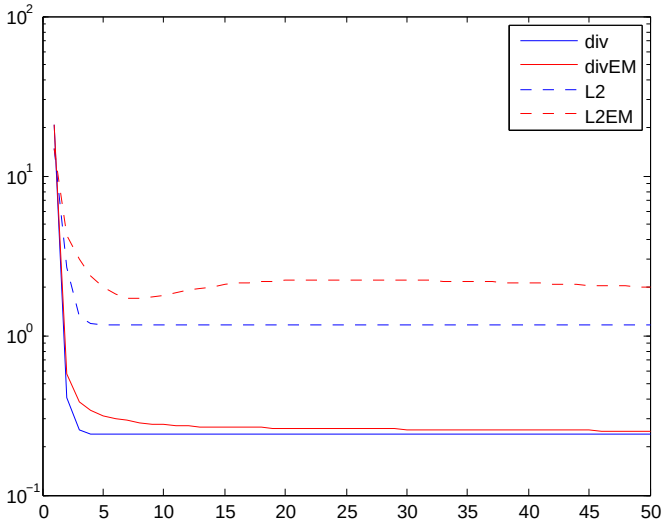

FIG 1. A=rand(10,5) , $d = 2*$rand(10,1) , k=2

completely fair to judge the quality of approximations by switching to another criterion. However, the L^2-norm of the error has an intuitive interpretation, is easy to compute and also has some appealing properties in the context of the two partial minimization problems, cf. Remarks 4.5 and 4.9.

We have plotted various characteristics of the algorithms against the number of iterations, for both of them the divergence at each iteration, as well as their counterparts for the L^2-norm (dashed lines). For reasons of clarity, in all figures we have displayed the characteristics on a logarithmic scale.

Legenda	
solid blue:	divergence $\mathcal{I}(\widehat{\Sigma}\|\|\widehat{\Sigma}_t)$ in algorithm 5.1
solid red:	divergence in the EM algorithm 6.1
dashed blue:	L^2-norm of $\widehat{\Sigma} - \widehat{\Sigma}_t$ in algorithm 5.1
dashed red:	L^2-norm in the EM algorithm 6.1

The numerical results have been obtained by running Matlab.

Figures 1 and 2 show the behaviour of the two algorithms in cases with $n = 10$ (which is relatively small) and for $k = 2, 5$ respectively. We observe that the performance of the algorithms hardly differ, especially for $k = 2$. In Figures 3 and 4, we have $n = 30$ and $k = 5, 15$ respectively. We notice that in terms of divergence, the performance of the two algorithms is roughly the same for $k = 5$, but for $k = 15$ there are noticeable differences. But looking

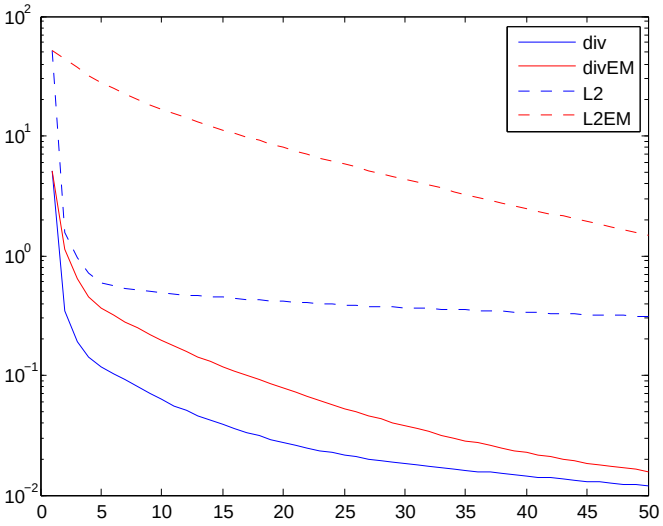

FIG 2. A=rand(10,5) , $d = 2*$rand(10,1) , k=5

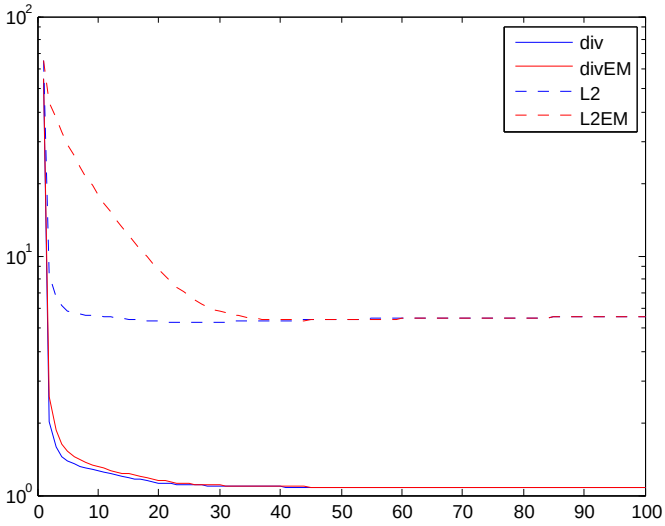

FIG 3. A=rand(30,15) , $d = 3*$rand(30,1) , k=5

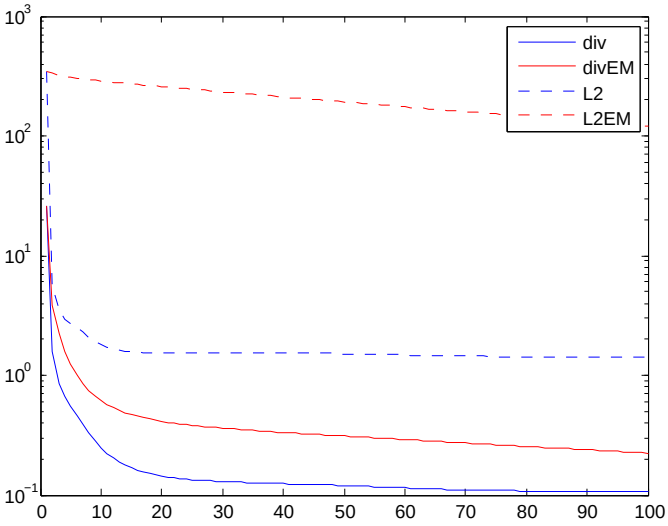

FIG 4. A=rand(30,15) , $d = 3*\text{rand}(30,1)$, k=15

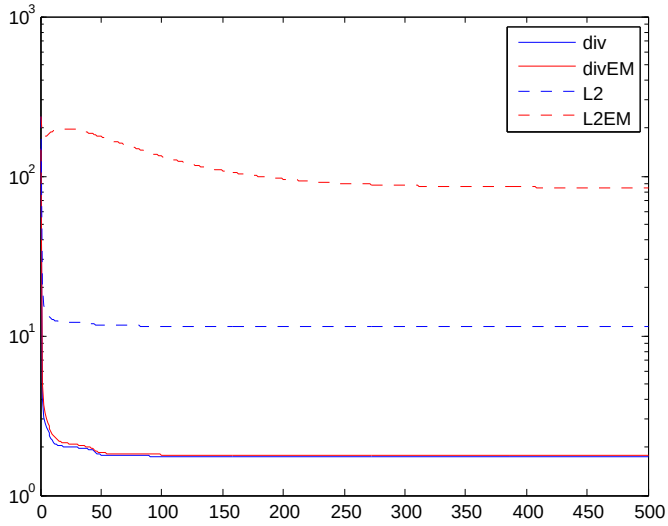

FIG 5. A=rand(50,30) , $d = 5*\text{rand}(50,1)$, k=10

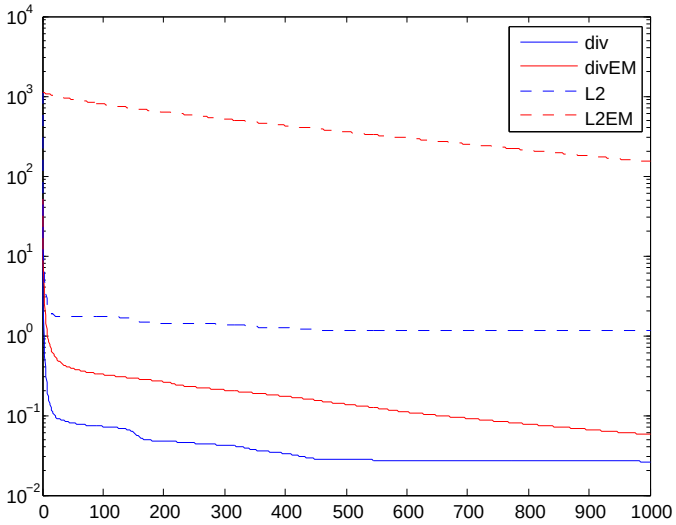

FIG 6. A=rand(50,30) , $d = 5*$rand(50,1) , k=30

at the L^2-norm of the error, we even see a manifest difference of the outcomes. The differences are even more pronounced in Figures 5 and 6, where $n = 50$ and $k = 10, 30$ respectively. In the former case, in terms of divergences, the two algorithms behave roughly the same, but there is a factor 10 of difference in the L^2-erros. In the latter case, where $k = m$ one would expect that both algorithms are able to retrieve the original matrix $\widehat{\Sigma}$, which seems to be the case, although Algorithm 5.1 behaves the best. Looking at the L^2-error, we see a gross difference between the Algorithm 5.1 and the EM algorithm of order about 100. This striking difference in behaviour between the two algorithms is typical.

8.2. Real data example

In the present section we test our algorithm on the data provided in the original paper Rubin and Thayer (1982), where the EM algorithm for FA models has been presented first. The results, with in this case $\widehat{\Sigma}$ the empirical correlation matrix of the data, are presented in Figure 7. We observe that again Algorithm 5.1 outperforms the EM algorithm. The underlying numerical results are at first sight very close to those of Rubin and Thayer (1982) (we have also taken $k = 4$), but we observe like in the previous section that the convergence of the EM algorithm is much slower than that of Algorithm 5.1 and after 50 iterations (the same number as in Rubin and Thayer (1982)) the differences are quite substantial.

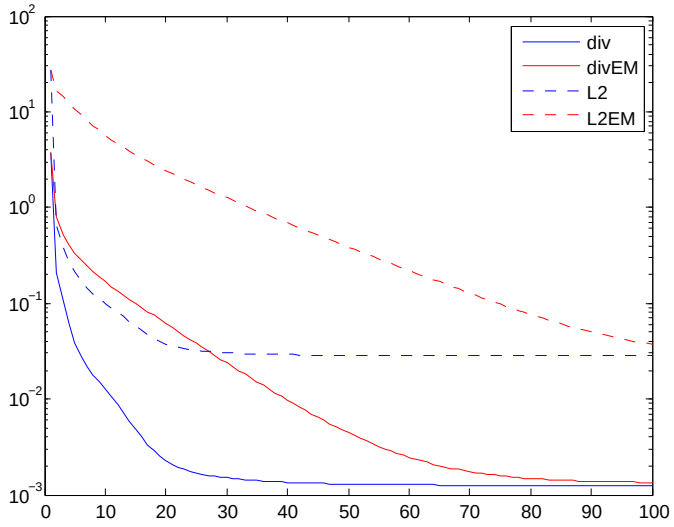

FIG 7. *Rubin-Thayer data, k=4*

Appendix A: Multivariate normal distribution

Let $(X^\top, Y^\top)^\top$ be a zero mean normally distributed random vector with covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}.$$

Assume that Σ_{YY} is invertible, then the conditional distribution of X given Y is normal, with mean vector $\mathbb{E}[X|Y] = \Sigma_{XY}\Sigma_{YY}^{-1}Y$ and covariance matrix

$$\mathbb{C}\mathrm{ov}[X|Y] = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}. \tag{A.1}$$

Consider two normal distributions $\nu_1 = N(\mu_1, \Sigma_1)$ and $\nu_2 = N(\mu_2, \Sigma_2)$ on a common Euclidean space. The I-divergence is easily computed as

$$\begin{aligned} \mathcal{I}(\nu_1||\nu_2) &= \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{m}{2} + \frac{1}{2}\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \frac{1}{2}(\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2) \\ &= \mathcal{I}(\Sigma_1||\Sigma_2) + \frac{1}{2}(\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2), \end{aligned} \tag{A.2}$$

where $\mathcal{I}(\Sigma_1||\Sigma_2)$ denotes as before the I-divergence between positive definite matrices. The extra term, depending on the nonzero means, did not appear in (3.2).

Appendix B: Matrix identities

For ease of reference we collect here some well known identities from matrix algebra.

The following lemma is verified by a straightforward computation.

Lemma B.1. *Let A, B, C, D be blocks of compatible sizes of a given matrix, with A and D both square. If D is invertible the following decomposition holds*

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} I & CD^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A - CD^{-1}B & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I & 0 \\ D^{-1}B & I \end{pmatrix}$$

while, if A is invertible, the following decomposition holds

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D - BA^{-1}C \end{pmatrix} \begin{pmatrix} I & A^{-1}C \\ 0 & I \end{pmatrix}.$$

Furthermore, assuming that A, D, and $A - CD^{-1}B$ are all invertible, we have

$$\begin{pmatrix} A & C \\ B & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - CD^{-1}B)^{-1} & -(A - CD^{-1}B)^{-1}CD^{-1} \\ -D^{-1}B(A - CD^{-1}B)^{-1} & D^{-1} + D^{-1}B(A - CD^{-1}B)^{-1}CD^{-1} \end{pmatrix}.$$

Corollary B.2. *Let A, B, C, D matrices as in Lemma B.1 with A, D, and $A - CD^{-1}B$ all invertible. Then $D - BAC$ is also invertible, with*

$$(D - BAC)^{-1} = D^{-1} + D^{-1}B(A^{-1} - CD^{-1}B)^{-1}CD^{-1}.$$

Proof. Use the two decompositions of Lemma B.1 with A replaced by A^{-1} and compute the two expressions of the lower right block of the inverse matrix. □

Corollary B.3. *Let $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{m\times n}$. Then $\det(I_n - BC) = \det(I_m - CB)$ and $I_n - BC$ is invertible if and only if $I_m - CB$ is invertible.*

Proof. Use the two decompositions of Lemma B.1 with $A = I_m$ and $D = I_n$ to compute the determinant of the block matrix. □

Corollary B.4. *Let D be a positive definite matrix, not necessarily strictly positive definite. If $HH^\top + D$ is strictly positive definite then also $I - H^\top(HH^\top + D)^{-1}H$ is strictly positive.*

Proof. Use Lemma B.1 with $A = I$, $B = H$, $C = H^\top$ and D replaced with $HH^\top + D$. The two middle matrices in the decompositions are respectively

$$\begin{pmatrix} I - H^\top(HH^\top + D)^{-1}H & 0 \\ 0 & HH^\top + D \end{pmatrix}$$

and

$$\begin{pmatrix} I & 0 \\ 0 & D \end{pmatrix}.$$

Hence, from the second decomposition it follows from positive definiteness of D that $\begin{pmatrix} I & H^\top \\ H & HH^\top + D \end{pmatrix}$ is positive definite, and then from the first decomposition that $I - H^\top(HH^\top + D)^{-1}H$ is positive definite. □

Appendix C: Decompositions of the I-divergence

We derive here a number of decomposition results for the I-divergence between two probability measures. Similar results are derived in Cramer (2000), see also Finesso and Spreij (2006) for the discrete case. These decompositions yield the core arguments for the proofs of the propositions in Sections 4.1 and 7.2.

Lemma C.1. *Let $\mathbb{P}_{XY}$ and $\mathbb{Q}_{XY}$ be given probability distributions of a Euclidean random vector (X, Y) and denote by $\mathbb{P}_{X|Y}$ and $\mathbb{Q}_{X|Y}$ the corresponding regular conditional distributions of X given Y. Assume that $\mathbb{P}_{XY} \ll \mathbb{Q}_{XY}$. Then*

$$\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}_Y||\mathbb{Q}_Y) + \mathbb{E}_{\mathbb{P}_Y}\mathcal{I}(\mathbb{P}_{X|Y}||\mathbb{Q}_{X|Y}). \tag{C.1}$$

Proof. It is easy to see that we also have $\mathbb{P}_Y \ll \mathbb{Q}_Y$. Moreover we also have absolute continuity of the conditional laws, in the sense that if 0 is a version of the conditional probability $\mathbb{Q}(X \in B|Y)$, then it is also a version of $\mathbb{P}(X \in B|Y)$. One can show that a conditional version of the Radon-Nikodym theorem applies and that a conditional Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}_{X|Y}}{\mathrm{d}\mathbb{Q}_{X|Y}}$ exists $\mathbb{Q}_Y$-almost surely. Moreover, one has the $\mathbb{Q}_{XY}$-a.s. factorization

$$\frac{\mathrm{d}\mathbb{P}_{XY}}{\mathrm{d}\mathbb{Q}_{XY}} = \frac{\mathrm{d}\mathbb{P}_{X|Y}}{\mathrm{d}\mathbb{Q}_{X|Y}}\frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{Q}_Y}.$$

Taking logarithms on both sides and expectation under $\mathbb{P}_{XY}$ yields

$$\mathbb{E}_{\mathbb{P}_{XY}} \log \frac{\mathrm{d}\mathbb{P}_{XY}}{\mathrm{d}\mathbb{Q}_{XY}} = \mathbb{E}_{\mathbb{P}_{XY}} \log \frac{\mathrm{d}\mathbb{P}_{X|Y}}{\mathrm{d}\mathbb{Q}_{X|Y}} + \mathbb{E}_{\mathbb{P}_{XY}} \log \frac{\mathrm{d}\mathbb{P}_Y}{\mathrm{d}\mathbb{Q}_Y}.$$

Writing the first term on the right hand side as $\mathbb{E}_{\mathbb{P}_{XY}}\{\mathbb{E}_{\mathbb{P}_{XY}}[\log \frac{\mathrm{d}\mathbb{P}_{X|Y}}{\mathrm{d}\mathbb{Q}_{X|Y}}|Y]\}$, we obtain $\mathbb{E}_{\mathbb{P}_Y}\{\mathbb{E}_{\mathbb{P}_{X|Y}}[\log \frac{\mathrm{d}\mathbb{P}_{X|Y}}{\mathrm{d}\mathbb{Q}_{X|Y}}|Y]\}$. The result follows. □

The decomposition of Lemma C.1 is useful when solving I-divergence minimization problems with marginal constraints, like the one considered below.

Proposition C.2. *Let $\mathbb{Q}_{XY}$ and $\mathbb{P}^0_Y$ be given probability distributions of a Euclidean random vector (X, Y), and of its subvector Y respectively. Consider the I-divergence minimization problem*

$$\min_{\mathbb{P}_{XY}\in\mathcal{P}} \mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}),$$

where

$$\mathcal{P} := \{\mathbb{P}_{XY} \mid \int \mathbb{P}_{XY}(dx, Y) = \mathbb{P}^0_Y\}.$$

If the marginal $\mathbb{P}^0_Y \ll \mathbb{Q}^0_Y$, then the I-divergence is minimized by $\mathbb{P}^*_{XY}$ specified by the Radon-Nikodym derivative

$$\frac{\mathrm{d}\mathbb{P}^*_{XY}}{\mathrm{d}\mathbb{Q}_{XY}} = \frac{\mathrm{d}\mathbb{P}^0_Y}{\mathrm{d}\mathbb{Q}_Y}. \tag{C.2}$$

Moreover the Pythagorean rule holds i.e., for any other distribution $\mathbb{P} \in \mathcal{P}$,

$$\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}_{XY}||\mathbb{P}^*_{XY}) + \mathcal{I}(\mathbb{P}^*_{XY}||\mathbb{Q}_{XY}), \tag{C.3}$$

and one also has

$$\mathcal{I}(\mathbb{P}^*_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}^0_{Y}||\mathbb{Q}_{Y}). \tag{C.4}$$

Proof. The starting point is equation (C.1), which now takes the form

$$\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}^0_{Y}||\mathbb{Q}_{Y}) + \mathbb{E}_{\mathbb{P}_Y}\mathcal{I}(\mathbb{P}_{X|Y}||\mathbb{Q}_{X|Y}). \tag{C.5}$$

Since the first term on the right hand side is fixed, the minimizing $\mathbb{P}^*_{XY}$ must satisfy $\mathbb{P}^*_{X|Y} = \mathbb{Q}_{X|Y}$. It follows that $\mathbb{P}^*_{XY} = \mathbb{P}^*_{X|Y}\mathbb{P}^0_Y = \mathbb{Q}_{X|Y}\mathbb{P}^0_Y$, thus verifying (C.2) and (C.4). We finally show that (C.3) holds.

$$\begin{aligned}
\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) &= \mathbb{E}_{\mathbb{P}_{XY}} \log \frac{\mathrm{d}\mathbb{P}_{XY}}{\mathrm{d}\mathbb{P}^*_{XY}} + \mathbb{E}_{\mathbb{P}_{XY}} \log \frac{\mathrm{d}\mathbb{P}^*_{XY}}{\mathrm{d}\mathbb{Q}_{XY}} \\
&= \mathcal{I}(\mathbb{P}_{XY}||\mathbb{P}^*_{XY}) + \mathbb{E}_{\mathbb{P}_Y} \log \frac{\mathrm{d}\mathbb{P}^0_Y}{\mathrm{d}\mathbb{Q}_Y} \\
&= \mathcal{I}(\mathbb{P}_{XY}||\mathbb{P}^*_{XY}) + \mathbb{E}_{\mathbb{P}^0_Y} \log \frac{\mathrm{d}\mathbb{P}^0_Y}{\mathrm{d}\mathbb{Q}_Y},
\end{aligned}$$

where we used that any $\mathbb{P}_{XY} \in \mathcal{P}$ has Y-marginal distribution $\mathbb{P}^0_Y$. □

The results above can be extended to the case where the random vector $(X, Y) := (X, Y_1, \dots Y_m)$, i.e. Y consists of m random subvectors Y_i. For any probability distribution $\mathbb{P}_{XY}$ on (X, Y), consider the conditional distributions $\mathbb{P}_{Y_i|X}$ and define the probability distribution $\widetilde{\mathbb{P}}_{XY}$ on (X, Y):

$$\widetilde{\mathbb{P}}_{XY} = \prod_i \mathbb{P}_{Y_i|X}\mathbb{P}_X.$$

Note that, under $\widetilde{\mathbb{P}}_{XY}$, the Y_i are conditionally independent given X. The following lemma sharpens Lemma C.1.

Lemma C.3. *Let* $\mathbb{P}_{XY}$ *and* $\mathbb{Q}_{XY}$ *be given probability distributions of a Euclidean random vector* $(X, Y) := (X, Y_1, \dots Y_m)$. *Assume that* $\mathbb{P}_{XY} \ll \mathbb{Q}_{XY}$ *and that, under* $\mathbb{Q}_{XY}$, *the subvectors* Y_i *of* Y *are conditionally independent given* X, *then*

$$\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}_{XY}||\widetilde{\mathbb{P}}_{XY}) + \sum_i \mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y_i|X}||\mathbb{Q}_{Y_i|X}) + \mathcal{I}(\mathbb{P}_X||\mathbb{Q}_X).$$

Proof. The proof runs along the same lines as the proof of Lemma C.1. We start from equation (C.1) with the roles of X and Y reversed. With the aid of $\widetilde{\mathbb{P}}_{XY}$

one can decompose the term $\mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X})$ as follows.

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y|X}||\mathbb{Q}_{Y|X}) &= \mathbb{E}_{\mathbb{P}_X}\mathbb{E}_{\mathbb{P}_{Y|X}}\log\frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\mathbb{Q}_{Y|X}} \\
&= \mathbb{E}_{\mathbb{P}_X}\mathbb{E}_{\mathbb{P}_{Y|X}}\left(\log\frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\widetilde{\mathbb{P}}_{Y|X}} + \log\frac{\mathrm{d}\widetilde{\mathbb{P}}_{Y|X}}{\mathrm{d}\mathbb{Q}_{Y|X}}\right) \\
&= \mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y|X}||\widetilde{\mathbb{P}}_{Y|X}) + \mathbb{E}_{\mathbb{P}_X}\mathbb{E}_{\mathbb{P}_{Y|X}}\sum_i \log\frac{\mathrm{d}\mathbb{P}_{Y_i|X}}{\mathrm{d}\mathbb{Q}_{Y_i|X}} \\
&= \mathcal{I}(\mathbb{P}_{XY}||\widetilde{\mathbb{P}}_{XY}) + \sum_i \mathbb{E}_{\mathbb{P}_X}\mathbb{E}_{\mathbb{P}_{Y_i|X}}\log\frac{\mathrm{d}\mathbb{P}_{Y_i|X}}{\mathrm{d}\mathbb{Q}_{Y_i|X}} \\
&= \mathcal{I}(\mathbb{P}_{XY}||\widetilde{\mathbb{P}}_{XY}) + \sum_i \mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y_i|X}||\mathbb{Q}_{Y_i|X}),
\end{aligned}$$

where we used the fact that $\frac{\mathrm{d}\mathbb{P}_{XY}}{\mathrm{d}\widetilde{\mathbb{P}}_{XY}} = \frac{\mathrm{d}\mathbb{P}_{Y|X}}{\mathrm{d}\widetilde{\mathbb{P}}_{Y|X}}$. This proves the lemma. □

The decomposition of Lemma C.3 is useful when solving I-divergence minimization problems with conditional independence constraints, like the one considered below.

Proposition C.4. *Let $\mathbb{P}_{XY}$ be a given probability distribution of a Euclidean random vector $(X,Y) := (X, Y_1, \dots Y_m)$. Consider the I-divergence minimization problem*

$$\min_{\mathbb{Q}_{XY}\in\mathcal{Q}} \mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}),$$

where

$$\mathcal{Q} := \{\mathbb{Q}_{XY} \mid \mathbb{Q}_{Y_1,\dots,Y_m|X} = \prod_i \mathbb{Q}_{Y_i|X}\}.$$

If $\mathbb{P}_{XY} \ll \mathbb{Q}_{XY}$ for some $\mathbb{Q}_{XY} \in \mathcal{Q}$ then the I-divergence is minimized by

$$\mathbb{Q}^*_{XY} = \widetilde{\mathbb{P}}_{XY}$$

Moreover, the Pythagorean rule holds, i.e. for any $\mathbb{Q}_{XY} \in \mathcal{Q}$,

$$\mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}_{XY}) = \mathcal{I}(\mathbb{P}_{XY}||\mathbb{Q}^*_{XY}) + \mathcal{I}(\mathbb{Q}^*_{XY}||\mathbb{Q}_{XY}).$$

Proof. From the right hand side of the identity in Lemma C.3 we see that the first I-divergence is not involved in the minimization, whereas the other two can be made equal to zero, by selecting $\mathbb{Q}_{Y_i|X} = \mathbb{P}_{Y_i|X}$ and $\mathbb{Q}_X = \mathbb{P}_X$. This shows that the minimizing $\mathbb{Q}^*_{XY}$ is equal to $\widetilde{\mathbb{P}}_{XY}$.
To prove the Pythagorean rule, we first observe that trivially

$$\mathcal{I}(\mathbb{P}_{XY}|\mathbb{Q}^*_{XY}) = \mathcal{I}(\mathbb{P}_{XY}|\widetilde{\mathbb{P}}_{XY}). \tag{C.6}$$

Next we apply the identity in Lemma C.3 with $\mathbb{Q}^*_{XY}$ replacing $\mathbb{P}_{XY}$. In this case the corresponding $\widetilde{\mathbb{Q}}^*_{XY}$ obviously equals $\mathbb{Q}^*_{XY}$ itself. Hence the identity reads

$$\begin{aligned}\mathcal{I}(\mathbb{Q}^*_{XY}||\mathbb{Q}_{XY}) &= \sum_i \mathbb{E}_{\mathbb{Q}^*_X}\mathcal{I}(\mathbb{Q}^*_{Y_i|X}||\mathbb{Q}_{Y_i|X}) + \mathcal{I}(\mathbb{Q}^*_X||\mathbb{Q}_X)\\ &= \sum_i \mathbb{E}_{\mathbb{P}_X}\mathcal{I}(\mathbb{P}_{Y_i|X}||\mathbb{Q}_{Y_i|X}) + \mathcal{I}(\mathbb{P}_X||\mathbb{Q}_X), \qquad \text{(C.7)}\end{aligned}$$

by definition of $\mathbb{Q}^*_{XY}$. Adding up equations (C.6) and (C.7) gives the result. □

Appendix D: Proof of the technical results

Proof of Proposition 3.2. *Existence of the minimum.* Let (H_0, D_0) be arbitrary. Perform one iteration of the algorithm to get (H_1, D_1) with $\mathcal{I}(\widehat{\Sigma}||H_1H_1^\top + D_1) \le \mathcal{I}(\widehat{\Sigma}||H_0H_0^\top + D_0)$. Moreover, from Proposition 5.4, $H_1H_1^\top \le \widehat{\Sigma}$ and $D_1 \le \Delta(\widehat{\Sigma})$. Hence the search for a minimum can be restricted to the set of matrices (H, D) satisfying $HH^\top \le \widehat{\Sigma}$ and $D \le \Delta(\widehat{\Sigma})$. We claim that the search for a minimum can be further restricted to the set of (H, D) such that $HH^\top + D \ge \varepsilon I$ for some sufficiently small $\varepsilon > 0$. Indeed, if the last inequality is violated, then $HH^\top + D$ has at least one eigenvalue less than ε. Assume this is the case, write $HH^\top + D = U\Lambda U^\top$, the Jordan decomposition of $HH^\top + D$, and let $\widehat{\Sigma} = U\Sigma_U U^\top$. Then $\mathcal{I}(\widehat{\Sigma}||HH^\top + D) = \mathcal{I}(\Sigma_U||\Lambda)$, as one easily verifies. Denoting by λ_i the eigenvalues of $HH^\top + D$, λ_{i_0} the smallest among them, and by σ_{ii} the diagonal elements of Σ_U, we have that $\mathcal{I}(\Sigma_U|\Lambda) = \frac{1}{2}\sum_i \left(\log\lambda_i + \frac{\sigma_{ii}}{\lambda_i}\right) - \frac{1}{2}\log|\Sigma_U| - \frac{n}{2}$. Choose ε smaller than $c := \min_i \sigma_{ii} > 0$, since $\widehat{\Sigma} > 0$. Then the contribution of $i = i_0$ in the summation is larger than $\log\varepsilon + \frac{c}{\varepsilon}$ which tends to infinity for $\varepsilon \to 0$. Hence the claim is verified. This shows that a minimizing pair (H, D) has to satisfy $HH^\top \le \widehat{\Sigma}$, $D \le \Delta(\widehat{\Sigma})$, and $HH^\top + D \ge \varepsilon I$, for some $\varepsilon > 0$. In other words we have to minimize the I-divergence over a compact set on which it is clearly continuous. This proves Proposition 3.2. □

Proof of Proposition 4.2. *Relation between the original and the lifted problem.* Let $\Sigma_1 = \Sigma(H, D, Q)$. With $\Sigma^* = \Sigma^*(\Sigma_1)$, the optimal solution of the partial minimization over $\mathbf{\Sigma}_0$, we have for any $\Sigma_0 \in \mathbf{\Sigma}_0$, using (4.3) in the first equality below,

$$\begin{aligned}\mathcal{I}(\Sigma_0||\Sigma_1) &\ge \mathcal{I}(\Sigma^*||\Sigma_1)\\ &= \mathcal{I}(\widehat{\Sigma}||HH^\top + D)\\ &\ge \inf_{H,D}\mathcal{I}(\widehat{\Sigma}||HH^\top + D).\end{aligned}$$

It follows that $\inf_{\Sigma_0\in\mathbf{\Sigma}_0,\Sigma_1\in\mathbf{\Sigma}_1}\mathcal{I}(\Sigma_0||\Sigma_1) \ge \min_{H,D}\mathcal{I}(\widehat{\Sigma}||HH^\top + D)$, since the minimum exists in view of Proposition 3.2.

Conversely, let (H^*, D^*) be the minimizer of $(H, D) \mapsto \mathcal{I}(\widehat{\Sigma}||HH^\top + D)$, which exists by Proposition 3.2, and let $\Sigma^* = \Sigma(H^*, D^*, Q^*)$ be a corresponding element in $\mathbf{\Sigma}_1$. Furthermore, let $\Sigma^{**} \in \mathbf{\Sigma}_0$ be the minimizer of $\Sigma_0 \mapsto \mathcal{I}(\Sigma_0||\Sigma^*)$ over $\mathbf{\Sigma}_0$. Then we have

$$\begin{aligned}\mathcal{I}(\widehat{\Sigma}||H^*{H^*}^\top + D^*) &= \mathcal{I}(\Sigma^{**}||\Sigma^*) \\ &\geq \inf_{\Sigma_0 \in \mathbf{\Sigma}_0, \Sigma_1 \in \mathbf{\Sigma}_1} \mathcal{I}(\Sigma_0||\Sigma_1),\end{aligned}$$

which shows the other inequality. Finally, we prove that the the infimum can be replaced with a minimum. Thereto we will explicitly construct a minimizer in terms of (H^*, D^*). For any invertible Q^* let $\Sigma^* = \Sigma(H^*, D^*, Q^*)$. Performing the first partial minimization, we obtain an optimal $\Sigma^{**} \in \mathbf{\Sigma}_0$, with the property (see (4.3)) that $\mathcal{I}(\Sigma^{**}|\Sigma^*) = \mathcal{I}(\widehat{\Sigma}||H^*{H^*}^\top + D^*)$. □

Proof of Proposition 4.4. *First partial minimization.* Consider the setup and the notation of Proposition C.2. Identify $\mathbb{Q}$ with the normal $N(0, \Sigma)$, and $\mathbb{P}$ with $N(0, \Sigma_0)$. By virtue of (C.2), the optimal $\mathbb{P}^*$ is a zero mean normal whose covariance matrix can be computed using the properties of conditional normal distributions (see appendix A). In particular

$$\begin{aligned}\Sigma^*_{21} = \mathbb{E}_{\mathbb{P}^*} XY^\top &= \mathbb{E}_{\mathbb{P}^*}(\mathbb{E}_{\mathbb{P}^*}[X|Y]Y^\top) \\ &= \mathbb{E}_{\mathbb{P}^*}(\mathbb{E}_{\mathbb{Q}}[X|Y]Y^\top) \\ &= \mathbb{E}_{\mathbb{P}^*}(\Sigma_{21}\Sigma_{11}^{-1}YY^\top) \\ &= \Sigma_{21}\Sigma_{11}^{-1}\mathbb{E}_{\mathbb{P}^0}YY^\top \\ &= \Sigma_{21}\Sigma_{11}^{-1}\widehat{\Sigma}.\end{aligned}$$

Likewise

$$\begin{aligned}\Sigma^*_{22} = \mathbb{E}_{\mathbb{P}^*} XX^\top &= \mathbb{C}\mathrm{ov}_{\mathbb{P}^*}(X|Y) + \mathbb{E}_{\mathbb{P}^*}(\mathbb{E}_{\mathbb{P}^*}[X|Y]\mathbb{E}_{\mathbb{P}^*}[X|Y]^\top) \\ &= \mathbb{C}\mathrm{ov}_{\mathbb{Q}}(X|Y) + \mathbb{E}_{\mathbb{P}^*}(\mathbb{E}_{\mathbb{Q}}[X|Y]\mathbb{E}_{\mathbb{Q}}[X|Y]^\top) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \mathbb{E}_{\mathbb{P}^*}(\Sigma_{21}\Sigma_{11}^{-1}Y(\Sigma_{21}\Sigma_{11}^{-1}Y)^\top) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \mathbb{E}_{\mathbb{P}^0}(\Sigma_{21}\Sigma_{11}^{-1}YY^\top\Sigma_{11}^{-1}\Sigma_{12}) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} + \Sigma_{21}\Sigma_{11}^{-1}\widehat{\Sigma}\Sigma_{11}^{-1}\Sigma_{12}.\end{aligned}$$

To prove that Σ^* is strictly positive note first that $\Sigma^*_{11} = \widehat{\Sigma} > 0$ by assumption. To conclude, since $\Sigma > 0$, it is enough to note that

$$\Sigma^*_{22} - \Sigma^*_{21}(\Sigma^*_{11})^{-1}\Sigma^*_{12} \;=\; \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

Finally, the relation $\mathcal{I}(\Sigma^*||\Sigma) = \mathcal{I}(\widehat{\Sigma}||\Sigma_{11})$ is Equation (C.4) adapted to the present situation. The Pythagorean rule follows from this relation and Equation (C.5). □

PROOF OF PROPOSITION 4.7. *Second partial minimization.* We adhere to the setting and the notation of Proposition C.4. Identify $\mathbb{P} = \mathbb{P}_{XY}$ with the normal distribution $N(0, \Sigma)$ and $\mathbb{Q} = \mathbb{Q}_{XY}$ with the normal $N(0, \Sigma_1)$, where $\Sigma_1 \in \mathbf{\Sigma}_1$. The optimal $\mathbb{Q}^* = \mathbb{Q}^*_{XY}$ is again normal and specified by its (conditional) mean and covariance matrix. Since $\mathbb{Q}^*_{Y_i|X} = \mathbb{P}_{Y_i|X}$ for all i, we have $\mathbb{E}_{\mathbb{Q}^*}[Y|X] = \mathbb{E}_{\mathbb{P}}[Y|X] = \Sigma_{12}\Sigma_{22}^{-1}X$, moreover $\mathbb{Q}^*_X = \mathbb{P}_X$. Hence we find

$$\Sigma^*_{12} = \mathbb{E}_{\mathbb{Q}^*} YX^\top = \mathbb{E}_{\mathbb{Q}^*}\mathbb{E}_{\mathbb{Q}^*}[Y|X]X^\top = \mathbb{E}_{\mathbb{P}}\mathbb{E}_{\mathbb{P}}[Y|X]X^\top = \Sigma_{12}.$$

Furthermore, under $\mathbb{Q}^*$, the Y_i are conditionally independent given X. Hence $\mathbb{C}\mathrm{ov}_{\mathbb{Q}^*}(Y_i, Y_j|X) = 0$, for $i \neq j$, whereas $\mathbb{V}\mathrm{ar}_{\mathbb{Q}^*}(Y_i|X) = \mathbb{V}\mathrm{ar}_{\mathbb{P}}(Y_i|X)$, which is the ii-element of $(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$, it follows that

$$\mathbb{C}\mathrm{ov}_{\mathbb{Q}^*}(Y|X) = \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

We can now evaluate

$$\begin{aligned}
\Sigma^*_{11} = \mathbb{C}\mathrm{ov}_{\mathbb{Q}^*}(Y) &= \mathbb{E}_{\mathbb{Q}^*} YY^\top \\
&= \mathbb{E}_{\mathbb{Q}^*}(\mathbb{E}_{\mathbb{Q}^*}[Y|X]\mathbb{E}[Y|X]^\top + \mathbb{C}\mathrm{ov}_{\mathbb{Q}^*}(Y|X)) \\
&= \mathbb{E}_{\mathbb{Q}^*}(\Sigma_{12}\Sigma_{22}^{-1}XX^\top\Sigma_{22}^{-1}\Sigma_{21} + \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})) \\
&= \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} + \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).
\end{aligned}$$

The Pythagorean rule follows from the general result of Proposition C.4. □

PROOF OF PROPOSITION 4.11. *Constrained second partial minimization.* Lemma C.3 and Proposition C.4 still apply, with the proviso that the marginal distribution of X is fixed at some $\mathbb{Q}^0_X$. The optimal distribution $\mathbb{Q}^*_{XY}$ will therefore take the form $\mathbb{Q}^*_{XY} = \prod_i \mathbb{P}_{Y_i|X}\mathbb{Q}^0_X$. Turning to the explicit computation of the optimal normal law, inspection of the proof of Proposition 4.7 reveals that under $\mathbb{Q}^*$ we have $\mathbb{E}_{\mathbb{Q}^*} YX^\top = \Sigma_{12}\Sigma_{22}^{-1}P_0$ and

$$\mathbb{C}\mathrm{ov}_{\mathbb{Q}^*}(Y) = \Delta(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) + \Sigma_{12}\Sigma_{22}^{-1}P_0\Sigma_{22}^{-1}\Sigma_{21}.$$

□

PROOF OF LEMMA 7.5. *Technical decomposition of the I-divergence.* Recall the following notation.

$$H = \begin{pmatrix} H_1 \\ H_2 \end{pmatrix} \tag{D.1}$$

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}, \tag{D.2}$$

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{21} & \widehat{\Sigma}_{22} \end{pmatrix}, \tag{D.3}$$

$$\widetilde{\Sigma}_{11} = \widehat{\Sigma}_{11} - \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{21}, \tag{D.4}$$

where $H_1 \in \mathbb{R}^{n_1 \times k}$, $H_2 \in \mathbb{R}^{n_2 \times k}$, $D_1 \in \mathbb{R}^{n_1 \times n_1}$ and $D_2 \in \mathbb{R}^{n_2 \times n_2}$.

Define $S = H_1(I - H_2^\top (H_2 H_2^\top + D_2)^{-1} H_2) H_1^\top + D_1$ and $K = \widehat{\Sigma}_{12}\widehat{\Sigma}_{22}^{-1} - H_1 H_2^\top (H_2 H_2^\top + D_2)^{-1}$. From Lemma C.1 we obtain that $\mathcal{I}(\widehat{\Sigma}||HH^\top + D)$ is the sum of $\mathcal{I}(\Sigma_{22}||H_2 H_2^\top + D_2)$ and an expected I-divergence between conditional distributions. The latter can be computed according to Equation (A.2), and gives the decomposition result

$$\mathcal{I}(\widehat{\Sigma}||HH^\top + D) = \mathcal{I}(\widehat{\Sigma}_{22}||H_2 H_2^\top + D_2) + \mathcal{I}(\widetilde{\Sigma}_{11}||S) + \frac{1}{2}\mathrm{tr}\{S^{-1} K \widehat{\Sigma}_{22} K^\top\}. \quad \text{(D.5)}$$

The assertion of Lemma 7.5 is then obtained by taking $D_2 = 0$ and the further decomposition of H_1 and H_2 as in (7.10). □

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* Wiley, New York.

Cramer, E. (2000). Probability measures with given marginals and conditionals: I-projections and conditional iterative proportional fitting. *Statistics and Decisions* **18** 311–329.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions* suppl. issue 1 205–237.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B* **39** 1–38.

Finesso, L. and Picci, G. (1984). Linear statistical models and stochastic realization theory. In *Analysis and optimization of systems* (A. Bensoussan and J. L. Lions, eds.). *Lecture Notes in Control and Information Sciences* 62 445–470. Springer, Berlin.

Finesso, L. and Spreij, P. (2006). Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications* **416** 270–287.

Finesso, L. and Spreij, P. (2007). Factor analysis and alternating minimization. In *Modeling, estimation and control, Festschrift in honor of Giorgio Picci* (A. Chiuso, S. Pinzoni and A. Ferrante, eds.). *Lecture Notes in Control and Information Sciences* 364 85–96. Springer, Berlin.

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** 443–482.

Rubin, D. B. and Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76.